INSTITUTO TECNOLÓGICO DE COSTA RICA

DEPARTAMENTO DE COMPUTACIÓN

PROGRAMA DE MAESTRÍA

Desarrollo de una metodología para la construcción de un Depósito de Datos, que combine e integre aspectos que representen y manipulen información imprecisa por medio de variables lingüísticas

Tesis para optar por el grado de Magíster Scientiae en Computación

Estudiantes:

Antonio Parra Arriola
Danilo Segura Murillo

Prof. Asesor:
Dr. Carlos González Alvarado

Cartago, Costa Rica
Junio 2001

Dedicatoria

A Dios, a la Virgen de los Ángeles y a mi Madre.
Con placer y muchísimo cariño dedico este esfuerzo que culmina con el presente documento a mi querida esposa, porque siempre supo comprenderme y tener paciencia durante el proceso de mis estudios, a mis hijos por ser un motivo más para seguir viviendo.

Danilo Segura Murillo

A Dios por darme la vida y brindarme la oportunidad de alcanzar esta meta.
A mi esposa porque siempre supo comprenderme y tenerme paciencia hasta en los peores momentos.
A mis hijos por ser la fuente inspiradora de mi superación y a mis padres por su amor y comprensión.

Antonio Parra Arriola

Agradecimientos

Nuestro profundo agradecimiento al Doctor Carlos González por su comprensión y oportunas sugerencias durante el desarrollo de esta investigación, así como su paciencia y vocación las veces que necesitamos de su asesoría. Su guía aseguró el éxito de esta empresa, siempre tuvo las observaciones necesarias para orientarnos cuando nuestro esfuerzo iba con rumbo equivocado.

Nuestra gratitud al MSc. Sergio Rodríguez Castillo por su desinteresada y entusiasta cooperación, su solidaridad profesional siempre será una fuente de inspiración. Los consejos, observaciones, comentarios y sugerencias fueron instrumental para llevar a cabo la conceptualización del tema.

A los compañeros de trabajo, que de muchas formas nos brindaron incondicionalmente su apoyo moral, técnico y espiritual.

A todas y a todos, nuestra eterna gratitud.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

La presión ejercida por la competencia en los negocios y el deseo de ser líder, ha impulsado a las organizaciones a explorar los beneficios de nuevas tecnologías que permitan ayudar a descubrir patrones de negocios en los datos, con miras a tener un mejor entendimiento del mercado y de sus clientes. Por este motivo, cualquier empresa que pretenda no quedar rezagada en su desarrollo debe estar al tanto de las técnicas que van surgiendo en el almacenamiento, transmisión y análisis de la información [SABA1995].

Tradicionalmente las bases de datos han sido las herramientas computacionales más utilizadas para llevar a cabo las tareas de almacenamiento y manipulación de grandes cantidades de datos. Con el objetivo de realizar las tareas eficientemente, estos sistemas cuentan con diversas técnicas para almacenar la información y facilitar el proceso de recuperación de aquella información que en un momento dado resulte necesaria, en un formato adecuado a las necesidades de la organización [GIRA1998].

A lo largo de los últimos años, han sido muchas y diferentes las aproximaciones que han surgido para cubrir todos estos objetivos, que se diferencian fundamentalmente en la forma y tipo de almacenamiento de los datos que son capaces de gestionar.

En la actualidad son pocas las empresas que pueden considerar como un tema de baja prioridad el estudio del impacto de la tecnología en su negocio. Este impacto puede asumir múltiples formas en una organización; desde obligarla a incorporar nuevas tecnologías para lograr un aumento en la eficacia, productividad y calidad, hasta situaciones en que la tecnología cambia totalmente el marco dentro del cual se mueve la empresa, la competencia y la forma de hacer negocios.

Ante esta exigencia las organizaciones han comprendido que las masas de datos almacenados contienen un importante e ignorado recurso. Un amplio conocimiento de sus negocios que explotado adecuadamente permitiría mejorar la gestión en la toma de

decisiones, por lo que han orientado sus esfuerzos a consolidar la información dispersa en un único repositorio que sirva de base para explotar el proceso de análisis de los datos. Estos elementos han permitido el inicio de una nueva actividad cuyo objetivo es hacer más eficiente los procesos de inferencia en masivos conjuntos de datos.

Como se ha mencionado, el acceso a la información es un elemento estratégico en la sociedad actual. Sin embargo, lo más importante no es la cantidad de información que se pueda acceder, si no la calidad de los mecanismos que se disponen para acceder a aquella información que nos interesa en un momento determinado.

En el mundo de los sistemas de apoyo a la toma de decisiones se han creado una gran variedad de arquitecturas, la más notable de estas estructuras es el depósito de datos (data warehouse) [INMO1995], [GIRA1998], [SABA1995], el cual contiene de forma integrada, resumida y detallada, datos históricos de la organización; específicamente estructurados para consulta y análisis.

En este trabajo se desarrolla un procedimiento innovador que describe la secuencia de tareas y actividades involucradas en la construcción de un depósito de datos, el cual servirá como herramienta de ayuda en el proceso de toma de decisiones, permitiendo ver nuevas tendencias y relaciones entre los clientes y los datos, también disponer de nuevas capacidades de análisis que se creían imposibles [IMMA1996], [SABA1995]. El elemento que ha permitido que los depósitos de datos tengan tanto auge hoy día es la importancia que dan las organizaciones a sus clientes, algunos de ellos hábitos de compra, volúmenes, quejas, y otros.

La representación de la información y el tratamiento de está se encuentran todavía lejos de los mecanismos de expresión utilizados habitualmente por el ser humano; en este sentido se está realizando un gran esfuerzo en el ámbito de la inteligencia artificial para resolver los problemas teóricos y prácticos relacionados con la elaboración de bases de datos más inteligentes.

Al existir una gran cantidad y variedad de datos cuya naturaleza no permite ser formulados de forma precisa, nuestra investigación pretende sentar las bases teórico-práctico para la implementación de un depósito de datos que combine e integre aspectos que representen y manipulen información expresada en términos imprecisos, en donde los aspectos más importantes de la información "imprecisa", que habitualmente manejamos son: la *incertidumbre* y la *imprecisión* en nuestras apreciaciones [KUYU1995]. El primero se deriva de apreciaciones realizadas sobre nuestra observación de la realidad que no pueden aportar un grado de certidumbre total sobre lo que afirmamos. El segundo aspecto se manifiesta a través del enunciado de conceptos que, o no se encuentran bien diferenciados o definidos, o bien resultan subjetivos. Para la representación y tratamiento de este tipo de información emplearemos variables lingüísticas, aparecidas en el campo de las bases de datos y de la teoría de conjuntos difusos de Zadeh [ZADE1965].

La presente investigación está dividida en cuatro capítulos, conclusiones y apéndices, organizados secuencialmente. El primer capítulo, presenta una perspectiva general del proceso de la toma de decisiones y cómo los sistemas han evolucionado, hasta llegar a lo que hoy conocemos como depósitos de datos [BAUM1996], [GIRA1998], [HAMM1996]. Además, en este capítulo se hace una descripción general de los principales elementos de un depósito de datos y su arquitectura, enfatizando los conceptos de Mercados de Datos (Data Marts) [INGL1997], INMO1996] y Modelo Multimensional de Bases de Datos [KIMB199], [KIMB1998], conceptos que serán empleados en el desarrollo de nuestro modelo. El segundo capítulo lo hemos dividido en tres partes, en la primera se presenta las características del "Modelo Relacional de Bases de Datos" [CODD1970], [CODD1990], [COYO1990], [DATE1990], [COSI1993], haciendo una descripción general de los elementos que configuran este modelo y enfatizando las limitaciones que presenta en el manejo de información imprecisa o incompleta [CODD1986], [KIRU1995]. En la segunda parte procedemos a plantear cómo la lógica difusa, y específicamente las variables lingüísticas [ZADE1995], permiten modelar los conceptos de información imprecisa, necesarios para la creación de un mercado de datos que incorpore datos imprecisos. En la tercera parte se desarrolla el modelo propuesto para el manejo de información imprecisa por medio de las variables

lingüísticas. En el tercer capítulo se desarrolla, desde el punto de vista teórico, la metodología propuesta para la construcción de un depósito de datos, describiendo para ello cada una de las actividades involucradas en las diferentes etapas. El cuarto capítulo se divide en dos partes, con el objetivo de aplicar los conceptos expuestos en los capítulos anteriores: en la primera parte se ilustra por medio de un caso práctico, la aplicación de las diferentes tareas involucradas en la metodología para la construcción de un depósito de datos; en la segunda parte se incorpora al depósito de datos los elementos difusos, utilizando para ello las diferentes fases de la metodología. Finalmente, se presentan las conclusiones y recomendaciones.

OBJETIVOS GENERALES

1. Desarrollar una metodología que permita la construcción de un depósito de datos y que además combine e integre aspectos que representen y manipulen información imprecisa, por medio de variables lingüísticas.
2. Construir un depósito de datos que reúna información de las colocaciones para una Institución Bancaria, aplicando para ello la metodología propuesta.

OBJETIVOS ESPECÍFICOS

1. Mostrar la utilidad del modelamiento multidimensional como una nueva forma de visualizar e interactuar con los datos.
2. Proponer una forma para representar la información imprecisa por medio de variables lingüísticas.
3. Demostrar, a través de un caso práctico, la utilidad de la metodología desarrollada.

CAPÍTULO 1

ESTADO DEL ARTE EN LA TOMA DE DECISIONES

"Todo lo que el hombre logra y todo lo que no puede lograr, es resultado directo de sus pensamientos."

James Allen

En este capítulo se presenta una perspectiva general del proceso de la toma de decisiones y de los fundamentos más relevantes de los Depósitos de Datos, Mercados de Datos y Modelamiento Multidimensional y cómo estos cambios tecnológicos han permitido que un número creciente de organizaciones pueda adoptar una nueva generación de infraestructura de la información que permita a los altos ejecutivos y analistas del negocio tener una mejor comprensión del mercado que sus competidores.

1.1 PERSPECTIVA DE LA TOMA DE DECISIONES

La necesidad de sistemas de información globales que permitan generar consultas, reportes y análisis para la toma de decisiones, es cada día más apremiante para las empresas que quieran competir exitosamente [GIRA1998].

Con el creciente y acelerado desarrollo de la tecnología informática y la popularización de los ambientes de cómputo distribuidos, la complejidad para recolectar, procesar y presentar la información se ha incrementado. Si a esto le sumamos la evolución y crecimiento de los mercados globalizados que provocan un incremento en las necesidades y requerimientos de información de ejecutivos y analistas, obtenemos como resultado una demanda, casi imposible de medir, de sistemas ejecutivos confiables y eficientes, lo que se traduce en un reto para las áreas de sistemas, que deben proveer estos servicios con infraestructura tecnológica y procesos sistematizados que garanticen el soporte para la toma de decisiones.

Desde hace muchos años las áreas de sistemas han venido trabajando para crear extractos de información de las bases de datos operacionales y almacenar estos datos en otra parte y formato, tratando de responder las peticiones de los usuarios por obtener información que les ayude a tomar mejores decisiones. Son pocas las organizaciones que han integrado exitosamente un sistema de información global para generar consultas, reportes y facilitar el análisis para la toma de decisiones.

Se ha acumulado una gran cantidad de datos a través de *sistemas transaccionales* que administran, por ejemplo; préstamos e inventarios, y procesos de contabilidad. En esta gran masa de datos existe información con el potencial de incrementar participaciones en el mercado, mejorar la productividad, aumentar el rendimiento de la inversión y mejorar el servicio al cliente. A pesar de lo anterior, las organizaciones se encuentran limitadas en su habilidad de hacer uso de dicha información en beneficio de la organización.

Los *sistemas operacionales* han sido construidos con el objetivo principal de dar soporte a las necesidades del día a día de la empresa [HAMM1996]. Son aplicaciones normalmente optimizadas para el manejo de un conjunto altamente estructurado y predefinido de transacciones, en donde cada una de las partes involucradas en el proceso utiliza el mismo grupo de datos, la diferencia son los objetivos, alcances, volúmenes y factores de tiempo que cada sector necesita. Como se muestra en la figura 1.1, lo que es común a todas es la diferencia en el uso de los datos y la perspectiva que el usuario tenga de los mismos.

Los *Sistemas de Información Ejecutiva* [GIRA1998] fueron soluciones populares en los 80’s, las que se caracterizaron por ser simples y poco flexibles, ya que el enlace entre datos y resultados reportados no permitían una exploración rápida y con el detalle necesario.

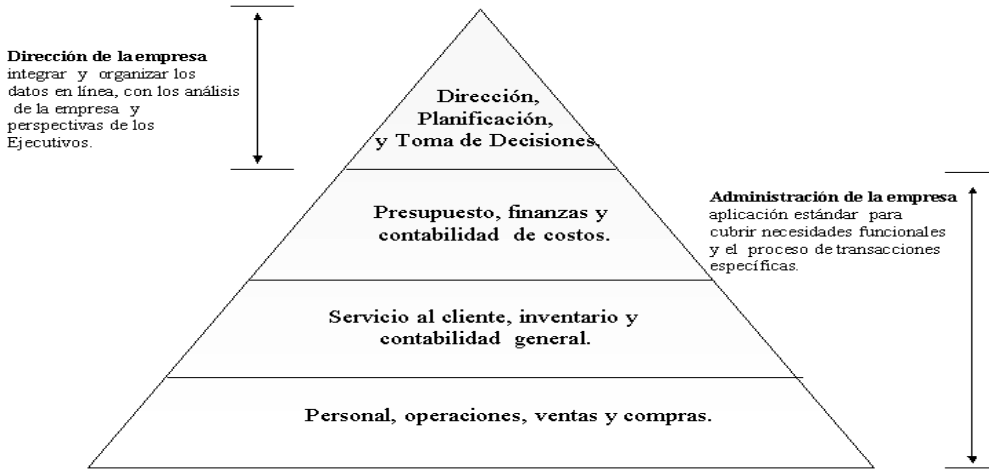


Figura 1.1 Uso de los datos en los diferentes niveles organizacionales

Para problemas específicos, los analistas desarrollaban aplicaciones a la medida, con herramientas de Sistemas de Soporte para las Decisiones o alternativamente, con herramientas de productividad personal (hojas electrónicas de cálculo), que debido a su capacidad limitada, alto grado de mantenimiento y sin posibilidad de compartirse, eventualmente fracasaba en su objetivo.

El término *depósito de datos* se ha propuesto como una solución a los problemas antes mencionados, este concepto será ampliado en el punto 1.2, página 10. Según S. Yazdani, [YAWO1998], la realidad de los negocios, y de nuestra vida en general, es que la era de la revolución industrial finalmente ha sido completada y que el mundo está entrando, en la era de la revolución de la tecnología y de la información, en donde los depósitos de datos son una de las manifestaciones de este nuevo período. Un depósito de datos está convirtiéndose más en una necesidad que un accesorio para las organizaciones que se consideren progresivas y competitivas.

Los avances en las computadoras y redes de comunicación han permitido la introducción de plataformas de hardware y software muy poderosos que permiten reunir, administrar y distribuir grandes cantidades de datos. Cualquier empresa que mantenga algún sistema de registro computadorizado y esté interesada en deducir o sacar

conclusiones lógicas de sus grandes volúmenes de información, debe considerar construir una aplicación de depósito de datos.

Estas empresas serán capaces de mejorar su perspicacia en el manejo de las tendencias de sus operaciones y eventualmente incrementar la exactitud de sus pronósticos y planes. La toma de decisiones tiene como base la información, la cual se formula por medio de datos obtenidos en experiencias previas, un depósito de datos almacena datos similares, apoyando de esta forma a los líderes en su proceso de toma de decisión.

La intuición, la "prueba y el error", ya no son métodos efectivos para administrar una empresa. La competencia se ha vuelto más agresiva, lo que ha obligado a los empresarios a estar cada vez más alerta sobre oportunidades que proporcionen un valor adicional a sus inversiones.

De acuerdo a T. Hammergren [HAMM1996] y S. Yazdani [YAWO1998], hay una serie de factores que pueden impactar negativamente en el proceso de toma de decisiones. Por ejemplo, algunos de los factores que afectan el análisis y las capacidades de investigación de los analistas de la empresa son:

- Crecimiento en el volumen de los datos.
- Los datos son almacenados en muchos diferentes sistemas y formatos.
- La introducción de nuevos productos.
- Dinamismo del mercado.
- Cambios en las estrategias de la organización.
- Lo crítico de tomar decisiones rápidamente.

Como resultado, muchos analistas utilizan poco de su tiempo en tareas analíticas, ya que la mayor parte de su tiempo se emplean en la extracción manual y preparación de los datos para investigación [YAWO1998]. Las tareas manuales reducen la eficiencia y efectividad de los procesos deductivos de análisis, lo que finalmente impacta en todo el

proceso de toma de decisiones. En un mundo con un mercado altamente competitivo, esta ineficiencia nos es aceptable.

Este problema se ve agravado por un modelo de empresa descentralizado, en donde la adquisición de las diferentes aplicaciones es también descentralizado. La falta de una adecuada planificación en el desarrollo de los sistemas transaccionales, ocasiona que estos sean islas de información que impiden a quienes tienen que tomar decisiones, ver de forma integrada los aspectos de cada situación y disponer de las variables necesarias para detectar problemas y oportunidades.

Estos sistemas han sido creados para acumular información en cada una de sus áreas de competencia, es decir, se han desarrollado verticalmente, especializándose en el área funcional a la cual pertenecen; en donde cada sistema maneja distintos formatos y convenciones para un mismo dato.

Los analistas de negocios, planificadores y ejecutivos, quienes se enfrentan con un mercado muy cambiante y una competencia muy dura, a menudo encuentran dificultades para plantear preguntas complejas de negocios a los diferentes sistemas. Esta deficiencia impacta su habilidad para responder a las tendencias del mercado rápidamente.

Con el paso de los años, las organizaciones siguen almacenando información y generando reportes, algunos con valor marginal. Hoy en día se ha hecho más evidente que se cuenta con una gran cantidad de datos, pero no de información. Debido al volumen y a la granularidad de los datos disponibles, los sistemas de información en producción sólo se pueden usar para apoyar la toma de decisiones operacional.

Los ejecutivos tienen el sentimiento de que no reciben la suficiente información útil de sus recursos de sistemas de información. Abogan por un sistema en el que se almacenen todas las operaciones diarias de la organización y adopte una perspectiva del negocio y permita el manejo de indicadores indistintamente de cómo los datos sean recolectados en cada una de las fuentes de información [GIRA1998], [KIMB1998]. Este

sistema debe ser rápido, flexible, extensivo, fiable y debe permitir al tomador de decisiones ver la organización de acuerdo a sus necesidades.

Hoy se vuelven cada vez más importantes los siguientes conceptos [HAMM1996], [BOAR1996]: la agregación y la sumarización. La *agregación* es la actividad que permite combinar varios conceptos, tales como las ventas de producto por cliente y por mercado. El concepto de *sumarización* permite la combinación de grandes cantidades de datos detallados para derivar resúmenes que tengan menos datos pero más información.

Es por esto, que los ejecutivos y el personal operativo recurren a la obtención de información agregada y condensada, lo que les permite comprender de forma más expedita las complejidades y peculiaridades de sus negocios, permitiendo visualizar de manera concisa las tendencias empresariales y de sus clientes, motivo principal de la implementación inicial de la tecnología de depósitos de datos.

Mientras los primeros implementadores luchaban contra la complejidad de implementar en forma manual extracciones de la base de datos, adición de datos, resúmenes de datos y análisis multidimensional de datos, otros intentaban usar Sistemas Ejecutivos de Información para obtener reportes resumidos y agregados. Sin embargo, su conclusión fue, por razones de desempeño, que era conveniente separar las bases de datos de producción, de las bases de datos del depósito de datos.

1.2 CONCEPTOS GENERALES SOBRE DEPÓSITOS DE DATOS

El término *Depósito de Datos* es un concepto de difusión relativamente reciente, un tanto amplio y en evolución, si se desea profundizar más sobre este concepto ver [ALUR1995], [BAUM1996], [YAWO1998], [INMO1995], [INMO1996], [INGL1997]. Esto hace que sea un poco difícil de categorizarlo adecuadamente y que, por

consiguiente, distintos proveedores de tecnología informática puedan adecuar el mensaje de venta de su producto a la palabra de moda.

Es una tecnología que tiene como objetivo resolver los problemas de manejo y uso adecuado de grandes fuentes de datos y de muy diversos tipos, para apoyar la toma de decisiones oportunas y fundamentadas [INGL1997]. Su arquitectura usualmente está basada en tecnología de Sistemas Administradores de Bases de Datos Relacionales (SABDR), usada para mantener datos que han sido extraídos del almacenamiento de datos operacionales [BOAR1996], [INGL1997]. Estos datos son transformados, consolidados y validados antes de ser cargados en él.

Literalmente las palabras que componen este término tienen las siguientes definiciones [HAMM1996];

- Dato: hecho e información acerca de algo.
- Depósito: lugar para almacenar bienes y mercaderías.

Un depósito de datos [INMO1995] provee un punto centralizado para almacenar los datos corporativos, y un conjunto de procesos consistentes y repetitivos para la carga de datos operacionales. Se construye sobre la base de una arquitectura escalable y abierta que permite manejar futuras expansiones de los datos. Además, provee un conjunto de herramientas que permiten a los usuarios procesar efectivamente los datos en información.

Como se ilustra en la figura 1.2, un depósito de datos es similar a una bodega tradicional de productos dentro de una industria de manufactura [HAMM1996]. Dentro del mundo de los sistemas de información es necesario adecuar los siguientes pasos.

1. Producir o comprar los materiales requeridos para elaborar el inventario, es decir, capturar los datos operacionales.

2. Tomar estos materiales y producir el inventario; o bien transformar los datos de las transacciones operacionales.
3. Almacenar el inventario final hasta que sea requerido por los canales de distribución; esto es, almacenar los datos transformados en el Depósito de Datos.
4. Distribuir el inventario de acuerdo a la demanda; en otras palabras satisfacer las consultas de los usuarios.

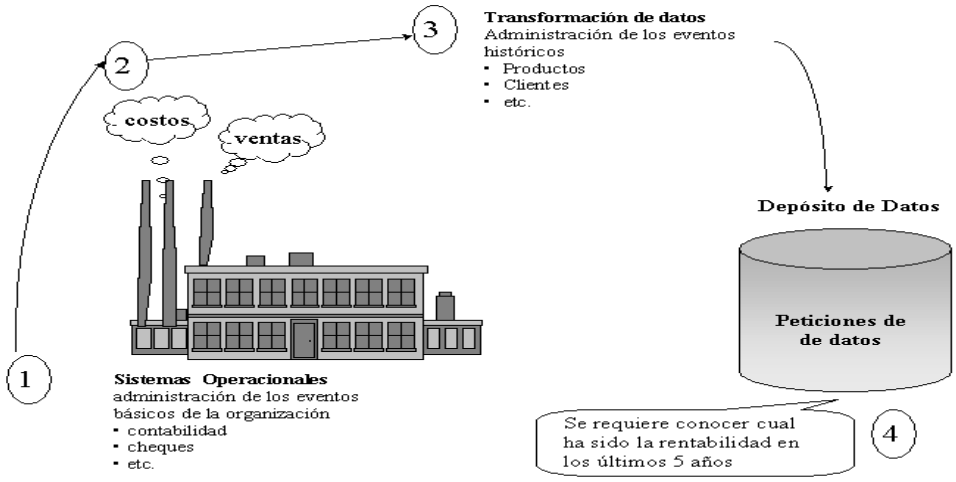


Figura 1.2 El proceso de generar un depósito de datos es análogo al proceso de manufactura

Algunos atribuyen el término depósito de datos a William Inmon [INMO1995], [INGL1997] Vicepresidente de PineCone System y reconocido como uno de los impulsores de esta arquitectura. De acuerdo con W. H. Inmon [INMO1995]: "Un depósito de datos es un conjunto de datos integrados orientados a una materia, que varían con el tiempo y que no son transitorios, los cuales soportan el proceso de toma de decisiones de una administración". Inmon lo define como aquel en que los datos son:

- *Orientados a asuntos*: La base de datos está orientada a asuntos, debido a que hay una transformación de datos que están orientados a aplicaciones de datos para el apoyo de toma de decisiones.

- *Integrados*: La base de datos es una combinación de varios sistemas, los cuales son diferentes entre sí.

- *Variables en el tiempo*: La información tiene una dimensión de tiempo; cada punto de datos esta asociado con un punto en el tiempo y estos puntos pueden ser comparados a lo largo del eje.

- *No son volátiles*: Siempre se están agregando nuevos datos a la base de datos en lugar de reemplazar los datos ya existentes.

Otros consideran a B. Devlin [DEVL1997] como el principal impulsor de este concepto. Devlin establece una clara división entre datos e información y define el depósito de datos como una arquitectura que permite extraer, depurar y consolidar datos de los sistemas operacionales para generar información que ayude en el proceso de toma de desiciones.

R. Kimball [KIMB1998] lo define como un depósito semánticamente consistente en datos (separados y que no interfieren con los sistemas operativos y de producción existentes) que llenan por completo los diferentes requerimientos de acceso y reporte de datos.

A su nivel más fundamental, un depósito de datos [DEVL1997],[KIMB1998], [INMO1995] es un área intermedia de información de apoyo al proceso de toma de decisiones. Recoge datos por medio de diversas aplicaciones en los sistemas operacionales de la organización, integra los datos en un modelo lógico en función del área de negocio, almacena la información en un formato que es accesible y comprensible, y distribuye dicha información por medio de diversas herramientas de consulta y de creación de informes para permitir la toma de decisiones de una forma ágil. Ver figura 1.3 en la que se representan gráficamente estos conceptos.

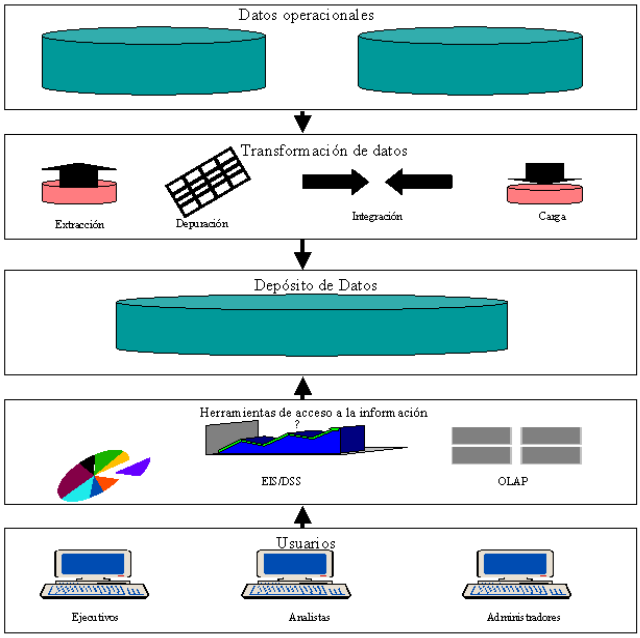


Figura 1.3 Componentes del proceso de un depósito de datos

Así como hay una gran divergencia para establecer una definición precisa de este concepto, hay un claro consenso de que esta tecnología es un ingrediente esencial en el conjunto de soluciones para el soporte de decisiones en una empresa.

Una definición de depósito de datos que nosotros consideramos adecuada es la siguientes: *"Es un proceso, no un producto, para consolidar y administrar datos de variadas fuentes con el propósito de responder a preguntas de negocios y tomar decisiones"*.

Ahora, ¿Qué es lo novedoso de esto? ¿No es lo que se ha estado pregonando desde hace un buen tiempo?. Lo novedoso es que los avances en la tecnología de hardware y software han desarrollado medios más adecuados para:

- *Consolidar datos desde una variedad de fuentes*. Dentro del marco conceptual de depósito de datos esta tarea se agrupa dentro del proceso de transformación de datos.

- *Manejar grandes volúmenes de datos de una forma que no era posible*. A estos medios los agruparemos en Procesamiento y Administración de Datos.

- *Acceder a los datos de una forma más directa, en el lenguaje del negocio, y analizarlos para obtener relaciones complejas entre los mismos*. El depósito de datos no podría ser aprovechado plenamente si no se hubieran elaborado herramientas que permitieran hacer un acceso inteligente al mismo.

Estos desarrollos tecnológicos, correctamente organizados e interrelacionados, constituyen lo que llamaremos un depósito de datos. Su aplicación puede tener variados fines, en una diversidad de industrias, no obstante, en términos generales, su utilización más rica corresponde a entornos de empresas en los que se identifican grandes volúmenes de datos, asociados a: cantidad de clientes, variedad de productos y cantidad de transacciones.

El máximo provecho de estos sistemas se obtiene a partir de la unificación de los datos de clientes y el almacenamiento de la información transaccional histórica. Esto permite realizar más eficientemente tareas de segmentación de clientes, identificación y análisis de mercados, cálculo de probabilidad de pérdida de clientes e implementación de campañas de retención, análisis de costos de adquisición de clientes y análisis de elasticidad-precio, entre otros.

El depósito de datos [GIRA1998], [HAMM1996], [KIMB1997] es un ambiente que puede proporcionar el poder y flexibilidad requerido por el negocio. Sin embargo, muchos de estos esfuerzos fallan, producto de que el modelo de datos no describe con exactitud el negocio y la plataforma tecnológica no es madura, firme, o probada y no es capaz de escalar significativamente por encima de las estimaciones realizadas.

Un ambiente de apoyo a la toma de decisiones puede consistir de un único depósito de datos en el cual se consoliden los datos corporativos para la toma de

decisiones, o de varios pequeños mercados de datos departamentales, o una combinación de ambos [INMO1999].

1.2.1 Sistemas Operacionales versus Depósitos de Datos

Antes de discutir más en detalle los conceptos de depósito de datos, es recomendable señalar algunas diferencias entre los sistemas operacionales y un sistema de depósito de datos.

Los objetivos de los sistemas operacionales y depósitos de datos para toma de decisiones son muy diferentes [HAMM1996]. Un *sistema operacional* asiste a la organización con la respuesta de los eventos o transacciones del día a día. Como resultado, este tipo de aplicaciones y sus datos son altamente estructurados alrededor de los eventos que ellos administran. Las bases de datos asociadas con estos sistemas son requeridas para soportar y procesar tan rápido como sea posible una gran cantidad de transacciones.

Los depósitos de datos generan bases de datos con una perspectiva histórica [HAMM1996], utilizando datos de múltiples fuentes que se fusionan en forma congruente y que permiten identificar oportunidades para incrementar sus ganancias, y de esta forma hacer que el negocio crezca. Estos datos se mantienen actualizados, pero no cambian al ritmo de los sistemas transaccionales.

Sin embargo, hay otras diferencias entre los dos sistemas, tales como las siguientes:

- *Tamaño y contenido*. Las metas y objetivos de un depósito de datos difieren ampliamente de las de un ambiente operacional. Un depósito de datos consiste básicamente de información histórica, la cual es necesaria para lograr un mejor entendimiento del negocio.

- *Rendimiento*. En un ambiente operacional la velocidad es esencial. Sin embargo, en un depósito de datos algunas consultas pueden tomar horas, lo cual puede ser aceptable, ya que la verdadera meta es proveer mejor información o inteligencia del negocio.

- *Vista parcial del negocio*. Los sistemas operacionales tienden a centrarse más en pequeñas áreas de la organización. Un depósito de datos permite ver todas las áreas funcionales de la institución.

- *Estructurado*. Los sistemas operacionales ofrecen pocas formas de accesar o entrar los datos que ellos administran. Un ambiente de depósito de datos se caracteriza por disponer de una gran variedad de mecanismos de visualización de la información.

- *Orientado a una materia*. Muchos sistemas operacionales organizan sus datos desde la perspectiva de la aplicación, de modo que el acceso de la aplicación a los datos tenga la mayor eficiencia posible. Con frecuencia, la información que está organizada para que una aplicación del negocio la recupere y actualice con facilidad, no está organizada necesariamente de modo que un analista con herramientas gráficas inteligentes de consulta pueda formular las preguntas empresariales correctas. Esto se debe al enfoque del diseño de la base de datos al momento en que se implemento por primera vez.

- *Nivel de detalle*. Con frecuencia es muy amplio el nivel de detalle de la información guardada por base de datos operacionales para cualquier toma de decisiones sensata. Un depósito de datos condensa y agrega la información para presentarla en forma comprensible a las personas. La condensación y adición es esencial para retroceder y entender la imagen global.

1.2.2 Arquitectura de un Depósito de Datos

Una arquitectura es un conjunto de reglas que proveen una estructura para el diseño global de un sistema o producto.

Según Tom Hammergren [HAMM1996], un depósito de datos no es ni un producto de software ni una máquina o tecnología de bases de datos en particular, sino una serie de componentes y procesos que en conjunto forman la arquitectura del depósito de datos. Sus componentes se muestran en la figura 1.4.

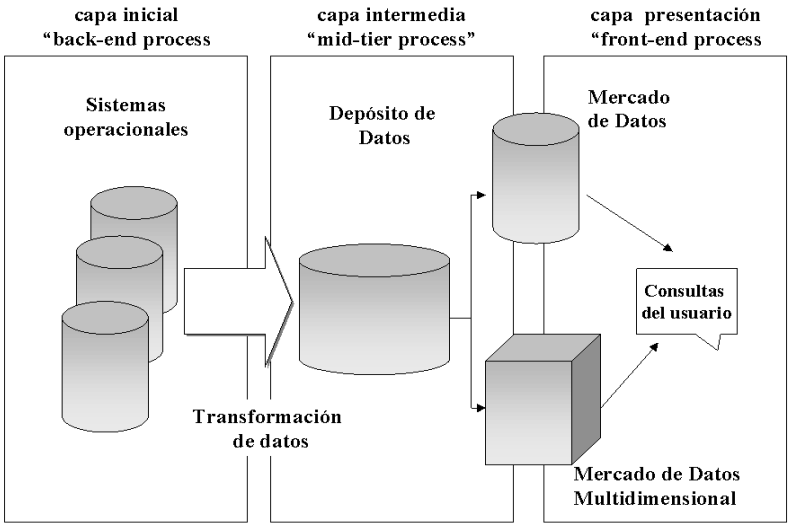


Figura 1.4 Arquitectura de un depósito de datos

La *capa inicial*, también llamada "*back-end process*" [HAMM1996], consiste en un conjunto de tareas de análisis, selección y extracción de datos que utilizan los datos almacenados en los sistemas operacionales. Estos procesos permiten extraer grandes volúmenes de datos de variadas fuentes a efectos de ser consolidados para su explotación.

Dichas tareas se instrumentan con herramientas de software que brindan una gran flexibilidad para acceder a distintas plataformas tecnológicas. Estos procesos son los más complejos en un sistema de depósito de datos, debido a que tienen que interactuar con múltiples fuentes, algunas de la cuales son fáciles de manipular, mientras que otras no.

Los datos deberán ser actualizados (extraídos de nuevo) como un proceso cíclico, periódico. Tener o pretender actualizaciones diarias es contraproducente en la mayoría de los casos, ya que la idea es trabajar con datos históricos con los cuales se puedan realizar comparaciones y proyecciones. Tratar de comparar datos de un día con el anterior no es una pregunta para toma de decisiones y en caso de serlo, no se necesita un Depósito de Datos para obtener la respuesta, ya que el sistema operacional sería suficiente.

La *capa intermedia* o "*mid-tier process*" [HAMM1996], está constituida por un grupo de tareas para homogeneizar y compatibilizar datos originalmente organizados de diferentes formas, esta es una característica importante que muchas veces se ignora o desestima. Si los datos provienen de fuentes diversas, con formatos diferentes, es necesario realizarles un proceso de transformación e integración, para convertirlos a un formato o formatos que permitan manejarlos en común, de acuerdo al modelo de datos de la empresa, y de acuerdo a la información para toma de decisiones con la que se desee contar.

A este nivel se cuenta con los detalles de los datos, así como con los metadatos o información de alto nivel que los describe. Para el almacenamiento y manejo de los datos extraídos se puede utilizar un Sistema Administrador de Bases de Datos Relacional, el cual, por su manejo de grandes volúmenes de información, por su implementación destinada a explotar las facilidades de hardware, los hace eficientes para consultas, selección y procesamiento de datos.

La *capa de presentación* o "*front-end process*" [HAMM1996] está constituida por las herramientas que permiten hacer un acceso inteligente al depósito de datos. Las mismas se han desarrollado de manera que las consultas puedan ser realizadas en forma sencilla, interactiva y en el lenguaje del negocio. La idea es que las preguntas sean realizadas por el usuario final – sea este un alto ejecutivo, un especialista de mercadeo, o un analista financiero – en su propio lenguaje, sin necesidad de conocer la tecnología subyacente, ni de hacer el requerimiento al área de sistema o a un grupo especializado.

Seguramente una sola herramienta no resolverá las necesidades de todas las posiciones dentro de la organización; pero existe una variedad tal que es posible cubrir los distintos requerimientos. Típicamente permiten un acceso orientado a temas, es decir: cliente, proveedor, producto, actividad, entre otros, y de una perspectiva histórica y geográfica. Es muy importante la interactividad, porque muchas respuestas a una pregunta de negocios sin duda sugerirán alguna otra pregunta relacionada que se deseará responder a efectos de la toma de decisiones.

Estas herramientas son las más sofisticadas y específicas de un depósito de datos, son las que más demanda ejercerá sobre la infraestructura del mismo y pondrán a prueba su diseño, especialmente en lo que respecta a la capacidad de procesamiento y administración de datos, a su vez son las que permiten obtener las relaciones más complejas y ocultas.

1.3 MERCADOS DE DATOS

Esta aproximación se creó como un complemento a la implantación de depósitos de datos, y su objetivo original es el crear pequeños depósitos satélites alrededor del depósito de datos corporativo [INMO1999], [MUNDI1997a], [MUNDI1997b], [STEV1997]. Los mercados de datos no solamente han sido utilizados como un complemento al depósito de datos central, sino como una estrategia sustituta, la cual permite crecer modularmente sin la necesidad de realizar grandes inversiones al inicio del proyecto, obteniendo resultados en tiempos relativamente cortos.

En el ambiente de depósitos de datos, un *mercado de datos* es un subconjunto de datos de propósito especial que son seleccionados para una función o aplicación particular [INMO1999]. Un mercado de datos está orientado a un grupo específico de usuarios mientras que un depósito de datos es construido para servir a toda la organización. Algunas veces este término se confunde con una base de datos departamental, en la cual los datos pertenecen únicamente a un departamento.

Un mercado de datos [INMO1999] es una forma sencilla de un depósito de datos, el cual está en función de una única área funcional tal como ventas, finanzas o mercadeo. Sin embargo, un mercado de datos es típicamente más pequeño y menos complejo que un depósito de datos y generalmente es más fácil de construir y administrar.

En un diseño apropiado de un depósito de datos, el marcado de datos es poblado y soportado a partir de una base de datos centralizada que contenga datos de toda la corporación. Un mercado de datos no debe ser creado fuera de la arquitectura de un depósito de datos [BAUM1996], [AMST1996], [INMO1999].

Los diferentes mercados de datos contienen diversas combinaciones y selecciones del mismo detalle de datos encontrados en el depósito. Sin embargo, los datos que residen en un depósito de datos tienen un nivel muy granular y los datos en el mercado de datos son más refinados.

En algunos casos el detalle de la información encontrada en el depósito de datos es localizada también en los diferentes mercados de datos. En otros casos, el detalle de información es diferente de un mercado de datos a otro. Pero en cada caso el depósito de datos es la fuente de todos los datos encontrados en los mercados de datos, todos tienen una herencia común. Estos se clasifican en mercados independientes o dependientes [INMO1999].

1.3.1 El mercado de datos independiente

Un *mercado de datos independiente* [BOAR1996], [INMO1999], es un subconjunto especifico de los datos operacionales orientados a un grupo determinado de usuarios. Son mercados organizados por área (tema) o grupos de usuarios, donde la información se obtiene directamente de los sistemas operacionales y se almacenan independientemente en un sistema físicamente separado. Esta solución es creada de forma independiente de una base de datos corporativa. En la figura 1.5 se muestra gráficamente este concepto.

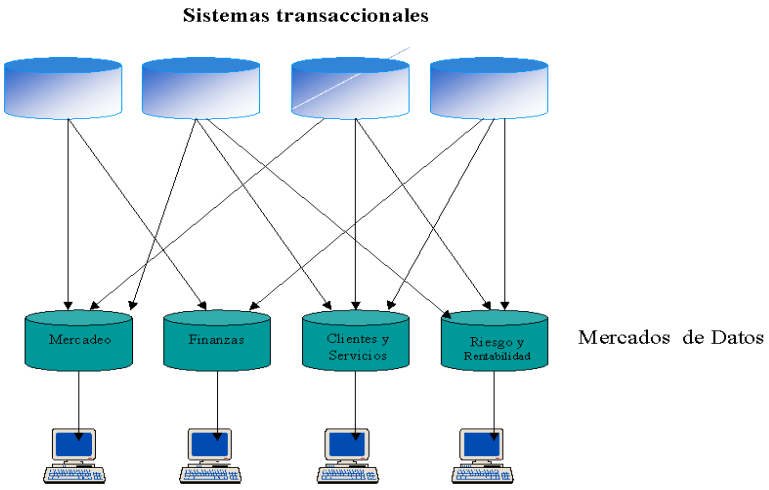


Figura 1.5 Mercados de datos independientes

Los mercados de datos independientes es la forma en que muchas organizaciones dan sus primeros pasos hacia un depósito de datos. Sin embargo, esta clase de soluciones únicamente soportan las necesidades de un departamento o un pequeño grupo de usuarios, no satisface las necesidades a nivel empresarial, en la cual hay muchas áreas funcionales y condiciones.

Las ventajas asociadas con un mercado de datos independiente son una rápida implementación y un control a nivel departamental. Su rápida implementación se debe a que se trabaja con un grupo de requerimientos y un modelo de datos pequeño, contrario a trabajar con un grupo de requerimientos y un modelo de datos a nivel empresarial. El control departamental es muy importante, dado que permite administrar la utilización del sistema y determinar cuando, donde y como se ejecutan las consultas.

Aun cuando esta clase de mercados pueden ser menos costosos de elaborar, su administración es muy onerosa. La duplicación de esfuerzos, tal como la transformación de los datos y el hecho de que se pueden encontrar datos incoherentes de un mercado a otro, es otra de las consideraciones que se deben analizar al construir un mercado de este tipo.

Cada mercado de datos independiente recibe datos de los sistemas operacionales. Por lo tanto, cada proceso de extracción y transformación ocurre muchas veces, una por cada mercado. Esto genera una duplicación de esfuerzos y un ineficiente uso de recursos.

Dado que los datos corporativos son almacenados en múltiples depósitos independientes, no se dispone de una vista global de los objetos de la empresa, tal como clientes, productos o regiones. Además, debido a que un mismo dato puede ser repetido a través de muchos mercados de datos, la redundancia es enorme.

Los modelos de datos de un mercado a otro pueden ser muy diferentes, esto hace que el proceso de integración sea muy difícil de lograr. Esta inconsistencia en la definición de los datos, es eliminada con un modelo empresarial, otra de las razones por las cuales se debe construir un depósito de datos.

1.3.2 El mercado de datos dependiente

Los *mercados de datos dependientes* son un subconjunto de datos de propósito especial que se obtienen a partir de un depósito de datos [BOAR1996], [INMO1999], en el cual los datos son seleccionados y organizados para un grupo particular de requerimientos o grupo de usuarios, y almacenados en un sistema de base de datos físicamente separado del depósito de datos. La fuente de los datos en un mercado de datos dependiente es el depósito de datos, en figura 1.6 se explica gráficamente este concepto.

En este esquema, semejante a una estrella, los datos son reorganizados en partes comunes que sean de utilidad a los usuarios, que datos son reunidos en cada mercado de datos depende de las necesidades de estos. Los usuarios tienen el segmento de datos que ellos necesitan en la forma en que ellos lo requieren.

Muchos proveedores abogan por la idea de construir mercados de datos dependientes para ocultar a los usuarios deficiencias de rendimiento. Un mercado de

datos dependiente debe ser construido basado en las necesidades de los usuarios y no en las limitaciones de rendimiento del depósito de datos.

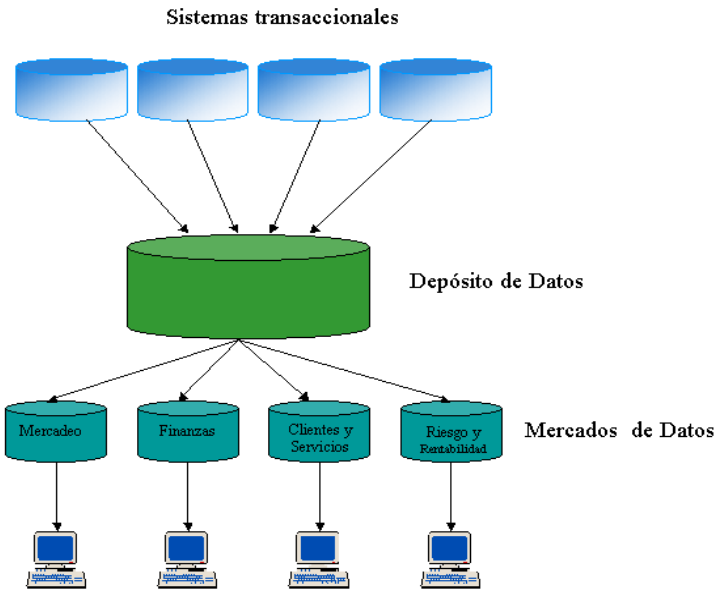


Figura 1.6 Mercados de datos dependientes

Con este tipo de soluciones se crea un ambiente en el cual los datos son vistos como un recurso corporativo a través del uso de un único modelo de datos; los datos son visualizados desde una perspectiva corporativa. Debido a que los datos en el depósito centralizado han pasado a través de un robusto proceso de transformación, el tiempo de implementación en un mercado de datos dependiente puede ser significativamente menor al tiempo requerido para construir un mercado de datos independiente.

Sin embargo, para disponer de una solución de esta clase, se requiere tener instalada toda la infraestructura que da soporte a un depósito de datos. Es importante entender cuando es el tiempo apropiado de construir un mercado de datos dependiente. Hay tres razones para construirlo:

- *Especificaciones de rendimiento*. Puesto que el subconjunto de datos es más pequeño que los datos contenidos en el depósito de datos, el tiempo de

consulta puede verse reducido. Consideraciones geográficas pueden tener un impacto en el rendimiento.

- *Control organizacional*. Debido a que puede haber una importante cantidad de presiones políticas dentro de la organización; conflictos entre las diferentes unidades del negocio, al construir mercados de datos dependientes cada departamento cree tener el control sobre los datos.

- *Aplicaciones especializadas*. Hay una gran cantidad de aplicaciones y herramientas que se especializan en proveer solución a un segmento especifico del negocio. Por ejemplo, análisis financiero, flujo de caja y análisis de riesgo en el área financiera/contable, segmentación de clientes y comercialización en el área de mercadeo.

1.3.3 ¿Por qué son atractivos los Mercados de Datos?

Hay muchas razones por las cuales el concepto de mercado de datos es atractivo. Su tiempo de implementación es más corto, aparentemente más fácil de administrar, y en teoría puede estar en producción más rápidamente y a un menor costo que los tradicionales depósitos de datos [ARAB1997], [AMST1996], [FRYE1996].

El argumento es que el mercado de datos es más pequeño. Es un conjunto conciso de información específica en un único lugar y por lo tanto más fácil de cargar y mantener. Este razonamiento es válido en aquellos ambientes en que se tiene únicamente un mercado de datos. Sin embargo, deja de ser cierto conforme más mercados de datos se van agregando.

Una segunda consideración es el uso de herramientas especializadas y muy poderosas que no tienen el grado de complejidad y el nivel de aprendizaje que se exige en ambientes de depósito de datos. Sin embargo, el principal problema con el ambiente de

los mercados de datos, es que son construidos para responder a situaciones como “se requiere el reporte XYZ”, más que el permitir análisis sobre los datos.

Los mercados de datos son una extensión natural de los depósitos de datos [INMO1999]. La cantidad de datos contenidos en él, está en función de las necesidades del departamento, y no de la corporación.

El costo de procesamiento y almacenamiento puede ser significativamente menor que el costo de procesamiento y almacenamiento en un ambiente de depósito de datos. Sin embargo, tiene el inconveniente de que no provee la capacidad de analizar como un todo, los datos de todas las áreas funcionales de la organización.

1.4 PROCESAMIENTO ANALÍTICO

Diariamente, los gerentes empresariales se enfrentan con dos retos fundamentales; operar la empresa de manera eficiente para maximizar la recuperación de la inversión y, planear el futuro. El procesamiento informático y el procesamiento analítico son dos formas básicas de aprovechar el depósito de datos para abordar estos dos retos [BAUM1996], [GIRA1998], [HAMM1996].

Con el procesamiento informático se buscan respuestas a cuestiones de operación tales como [GIRA1998]; ¿Cuáles fueron los ingresos por ventas? ¿Cuáles fueron los diez productos más rentable durante el último periodo de ventas?. Los mismos gerentes y analistas requieren la funcionalidad del procesamiento analítico cuando se deben responder preguntas complejas como ¿Cuántas tarjetas de crédito se vendieron a hombres en el mes de diciembre, en nuestras oficinas de la región norte? ¿Cómo se compara lo programado con lo real del mismo mes en los últimos dos años?.

Los gerentes, ejecutivos y analistas del negocio saben que el futuro pertenece a quienes pueden verlo y llegar ahí primero. Por tal razón los ejecutivos y gerentes no sólo comprenden lo que está pasando en el negocio, sino también que va a suceder.

Desde la perspectiva de análisis, los usuarios empresariales desean extraer los datos adecuados de manera intuitiva, con una mínima inversión de tiempo. Una vez que extraen los datos correctos del depósito de datos, los analizan, sintetizan y consolidan en información, para después emplear esa información a fin de tomar mejores decisiones.

En el marco de referencia de la plataforma de soporte de decisiones, el procesamiento analítico [GIRA1998], se circunscribe principalmente a la modalidad de uso de verificación. En esta modalidad, el usuario empresarial crea una hipótesis y accede a datos para verificarla o confirmarla. Por lo general el análisis de los datos es iterativo, una cuestión empresarial conduce a otra, hasta que se deduce un conjunto claro de alternativas y recomendaciones de acción potenciales. Durante esta actividad de análisis, los usuarios descubren en ocasiones relaciones insospechadas entre parámetros empresariales, por ejemplo, la relación de las ventas con la edad y el sexo de los clientes.

1.4.1 Análisis multidimensional

Los datos empresariales son, de hecho, multidimensionales [KIMB1997], [HAMM1996], [KIMB1998]. Se encuentran relacionados y regularmente son jerárquicos; por ejemplo, los datos de ventas y los pronósticos de presupuesto están interrelacionados y dependen entre sí. En la práctica, para predecir las ventas de un nuevo producto, se requiere analizar los patrones de compras anteriores, la adopción de nuevos productos, las preferencias y otros factores similares.

Mejorar la satisfacción del cliente al mismo tiempo que se conserva un margen de rentabilidad competitivo, es un reto monumental. Tal reto requiere entender un conjunto complejo de dimensiones empresariales interrelacionadas, por lo que resulta fundamental

el análisis e interpretación de los datos desde varias perspectivas [GIRA1998], [KIMB1997]. Para lograr esto es un requisito el análisis multidimensional.

En el análisis multidimensional, los datos se representan mediante dimensiones como producto, territorio y cliente [KIMB1997], [MENN1996], [PARS1995], un ejemplo de esto se muestra en la figura 1.7. Por lo regular las dimensiones se relacionan en jerarquías, por ejemplo, país, región, provincia, cantón y distrito. El tiempo también es una dimensión estándar con su propia jerarquía como día, semana, mes, trimestre y año.

Para facilitar un análisis complejo, el análisis multidimensional presenta una visión empresarial sencilla de los datos. Un usuario puede acceder los ingresos por área y departamento para los últimos cuatro trimestres, para un conjunto dado de productos. Los resultados se pueden pivotar o girar para cambiar los ejes y la perspectiva. Además, los usuarios pueden navegar por las dimensiones obteniendo de esta forma resúmenes a lo largo de los elementos de una dimensión, o penetrar a través de las dimensiones para ver otras perspectivas

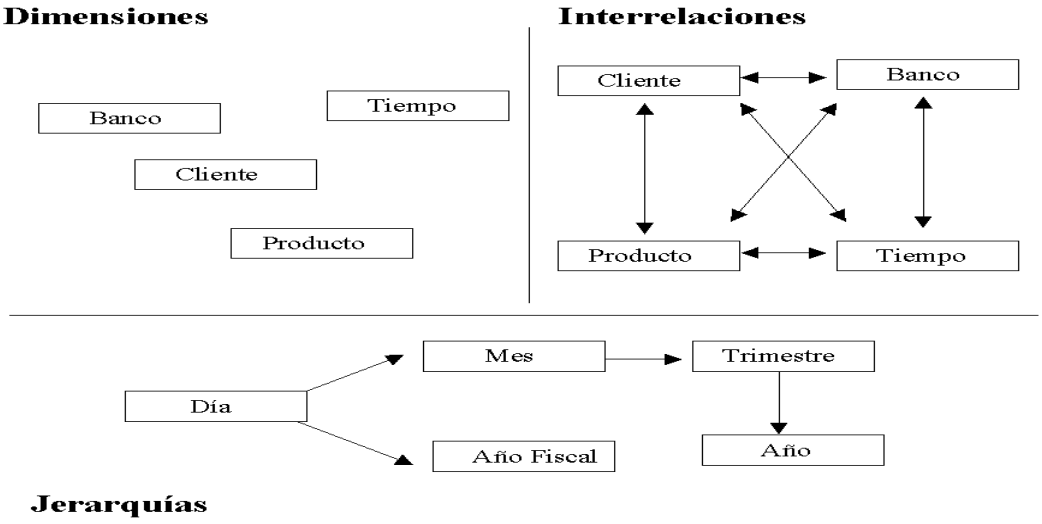


Figura 1.7 Análisis multidimensional

El procesamiento analítico se utiliza para entender lo que está sucediendo en la empresa y promete análisis del tipo “¿qué pasa si...?” o “¿y ahora qué?”. Además, se

emplea para análisis históricos complejos, con amplia manipulación (análisis de datos dinámicos), así como para planeación a futuro y pronósticos[GIRA1998]. El procesamiento informático es por lo regular un análisis más sencillo (de dos o tres dimensiones) de datos históricos para comprender el pasado (análisis de datos estáticos).

El procesamiento analítico o análisis multidimensional se le conoce también como procesamiento analítico en línea (OLAP). Se apoya en una visión multidimensional de los datos en el Depósito de Datos.

1.4.2 Procesamiento analítico en línea (OLAP)

Como cualquier otra forma de modelaje semántico, los modelos relacionales tradicionales se centran en aspectos estructurales, algunas veces en detrimento de los aspectos de manipulación de la información [BAUM1996], [INGL1997], [KIMB1998]. La estructura relacional consiste de tablas y relaciones, mientras que la visión del usuario consiste de jerarquías y dimensiones, que le permiten observar el negocio desde diferentes perspectivas [KIMB1997], [THOM1997].

Los sistemas de procesamiento de transacciones formalmente conocidos como sistemas de procesamiento transaccional en línea (OLTP), son muy eficientes operando sobre información detallada y no redundante. Paralelamente, este tipo de sistemas es en esencia distinta de aquellos sistemas que podemos denominar “de toma de decisiones”.

Los sistemas OLTP [GIRA1998] actúan sobre datos actuales, reales y que corresponden con sistemas operacionales de la empresa, las operaciones se apoyan sobre mecanismos de optimización particularizados para una cierta clase de transacciones predeterminadas y finalmente las transacciones en esta clase de aplicaciones actúan sobre un pequeño conjunto de datos en una operación de muy corta duración. Por su parte, los sistemas de toma de decisiones [GIRA1998] requieren analizar datos históricos, de una manera flexible y sobre información resumida o sumarizada.

Por esta discrepancia de fondo, algunos autores [GIRA1998], [WHIP1997], [THOM1997], [WILL1998] han comenzado a formular un concepto que diferencie los sistemas transaccionales en línea de los sistemas de toma de decisiones. Este nuevo concepto se conoce como procesamiento analítico en línea (OLAP).

El procesamiento analítico en línea es una tecnología de análisis de datos que hace lo siguiente:

- Presenta una visón multidimensional lógica de los datos en el depósito de datos. La visión es independiente de cómo se almacenan los datos.

- Maneja modelos funcionales de pronóstico, análisis de tendencias y análisis estadísticos.

- Tiene un motor de depósito de datos multidimensional, que almacena los datos en arreglos. Estos arreglos son una representación lógica de las dimensiones organizacionales.

- Responde con rapidez a las consultas, de modo que el proceso de análisis no se interrumpe y la información no se desactualiza.

- Ofrece opciones de modelado analítico, incluyendo un motor de cálculo para obtener proporciones, desviaciones, entre otros, que comprende mediciones de datos numéricos a través de muchas dimensiones.

- Crea resúmenes y adiciones, jerarquías, y cuestiona todos los niveles de adición y resumen en cada intersección de las dimensiones.

La tecnología OLAP se aplica en muchas áreas funcionales de una organización, tales como ventas, análisis de rentabilidad, consolidaciones financieras, presupuestos y

pronósticos y contabilidad de costos. Es una opción de análisis y de reporte, es un componente importante de la arquitectura de un depósito de datos.

En esencia OLAP [GIRA1998] es una base de datos y un conjunto de herramientas de análisis que permiten a los usuarios finales observar y analizar información gerencial rápidamente, de una manera natural y flexible. La información sobre la cual se opera en un sistema de esta clase proviene de las aplicaciones operacionales de la empresa, que probablemente se encuentran implementadas bajo modelos relacionales.

La arquitectura funcional OLAP consta de tres componentes de servicio: a) OLAP b) de depósitos de datos, y c) de presentación a usuarios. A este respecto la arquitectura funcional es una arquitectura Cliente-Servidor que ofrece varias opciones de configuración física para los tres servicios funcionales. La arquitectura física consta de dos grandes categorías basadas en tecnologías de depósitos de datos: depósito de datos multidimensional y depósito de datos relacional.

Desde una perspectiva de arquitectura global no es fácil elegir entre el depósito de datos multidimensional o el relacional para OLAP. La organización necesita proporcionar los criterios para hacer la selección adecuada. Sin embargo, la tendencia de la industria es ofrecer los servicios OLAP con una combinación de un proceso frontal de servidor OLAP (con un depósito multidimensional incorporado para datos grueso) y un proceso posterior de depósito relacional (con un nivel fino de datos detallados).

El éxito en las empresas que puedan implementar aplicaciones OLAP, dependerá frecuentemente en sus experiencias y la metodología empleada. Puesto que las aplicaciones de OLAP son principalmente desarrolladas y mantenidas por el usuario, no requieren de un soporte amplio por las áreas de sistemas. La innovación en el proceso de la toma de decisiones se basa en facilitar un entorno mediante herramientas analíticas bien probadas y de técnicas que ayudan a determinar perspectivas claras sobre la dinámica del negocio.

Las aplicaciones OLAP brindan la inteligencia de negocios requerida para tomar decisiones, y por consiguiente definen la estructura y contenido del depósito de datos, el nivel histórico y de detalle requerido de la información.

1.4.3 Bases de datos multidimensionales

Las bases de datos multidimensionales han existido desde hace algunos años. Aquí, la información se almacena dimensional y no relacionalmente [KIMB1997], [DELG1997]. Las dimensiones determinan la estructura de la información almacenada y definen adicionalmente caminos de consolidación. La información almacenada se presenta como variables que a su vez están caracterizadas por una o más dimensiones. De este modo, la información puede analizarse dentro del cubo formado por la intersección de las dimensiones de la variable particular. Por ejemplo, el gerente de producto podría estar interesado en analizar la información en el plano horizontal, el cual representa el producto por el cual es responsable; visión del producto. Un ejemplo de esto se muestra en la figura 1.8.

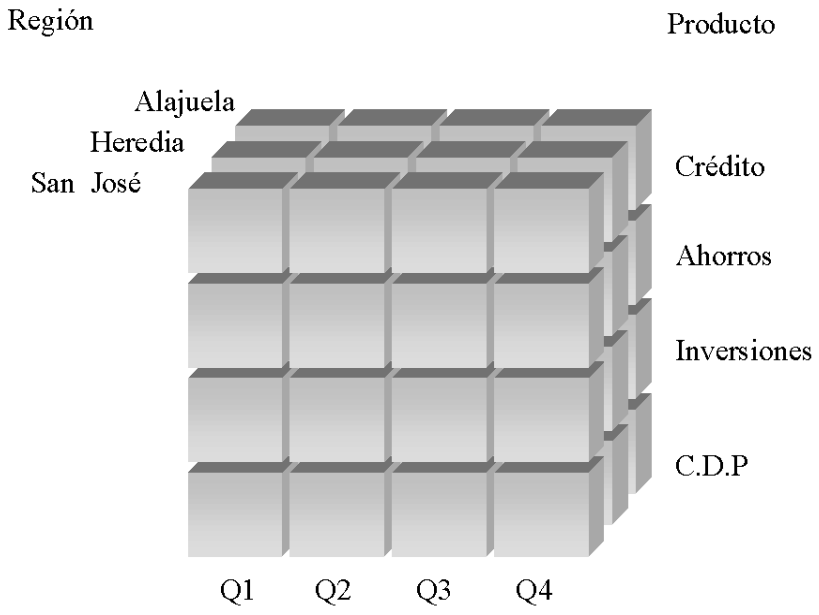


Figura 1.8 Cubo multidimensional

De una manera similar, el gerente de área puede seccionar el cubo en un plano vertical que representa la región a su cargo; visión regional. En todo momento, el gerente financiero puede comparar la información de dos años diferentes o bien hacer análisis de tendencias de las ventas, mediante el corte del cubo en su profundidad; visión temporal.

A su vez, las dimensiones pueden estructurarse jerárquicamente, de modo que se pueden construir caminos de consolidación. Estas rutas permiten analizar la información bajo un mecanismo “drill-down”, de lo más general hacia lo más específico.

En general, el análisis sobre bases de datos multidimensionales, conocido como MOLAP [GIRA1998],[DELG1997] en contraposición con ROLAP o OLAP relacional, provee capacidades de:

- *Análisis comparativo o relativo*: ¿Cómo las ventas actuales se comportan con respecto a las ventas esperadas?

- *Reporte de excepciones o tendencias*: ¿Cuáles productos se han vendido menos del 5% de lo esperado y representan más del 2% de las ventas totales?

- *Modelaje, proyecciones*: ¿Qué pasaría si se agregan tres productos más a la región central?

En el esquema MOLAP [GIRA1998], [DELG1997], los datos se almacenan de manera lógica en arreglos, utilizando uno de dos enfoques: el hipercubo o el multicubo. En el enfoque de hipercubo, los objetos con tres o más dimensiones se describen como lados planos y cada dimensión con ángulos rectos en relación con los demás.

Los datos se organizan de acuerdo con la visión empresarial de los mismos y, como regla se almacenan en forma resumida o agregada. El índice es más pequeño, dando por resultado una respuesta muy rápida a consultas complejas. Debido a que los valores se almacenan en arreglos, la actualización del valor no afecta el índice. Esta

característica facilita la implementación de aplicaciones de lectura-escritura o de actualización, como la de pronósticos y presupuestación.

Los SABDR y sus herramientas asociadas, pueden ser limitadas cuando se requieren tareas de análisis, debido a que se cuenta con un manejo ineficiente de relaciones multidimensionales, y poca habilidad de análisis y consolidación. Esto se debe principalmente a que el modelo relacional se basa en un esquema bidimensional (tablas) y el manejo de datos mediante "SQL" no es adecuado para los cálculos que requiere una aplicación OLAP [KIMB1997], [KIMB1998], [THOM1997].

Los requerimientos de un sistema OLAP y el deseo de aventajar con las bases de datos relacionales, ha conducido a la creación de productos referidos como OLAP relacionales (ROLAP). Estas herramientas, extienden la capacidad de las bases de datos relacionales en un intento de hacerlas también adecuadas para aplicaciones OLAP.

Las herramientas ROLAP [GIRA1998], [DELG1997] mapean una estructura multidimensional en una capa por encima de la estructura de tablas original. Aunque los datos se almacenan en forma relacional (columnas y filas), se presentan al usuario en forma de dimensiones. Para lograrlo se crea una capa semántica de metadatos que permite ubicar las dimensiones para las tablas relacionales. También se crean metadatos adicionales para cualquier resumen o adición, con el fin de mejorar el tiempo de respuesta

El análisis ROLAP, a pesar de ser más sencillo de construir y más fácil de mantener, presenta algunas desventajas:

- La mayoría de las necesidades de análisis requieren que la información sea procesada en un modelo de series de tiempo. En un sistema relacional, donde el lenguaje de acceso es "SQL", preguntas como ¿cuánto han variado las ventas con respecto al promedio del último año?, son extremadamente difíciles de responder.

- Debido a que la base de datos operacional, se encuentra altamente estructurada, un cambio en los requerimientos, o la inclusión de una nueva variable para el análisis, representa un cambio mayor en el modelo de la base de datos.

- El tiempo para construir un modelo multidimensional basado en una estructura relacional de la información, con el objetivo de resolver los inconvenientes anteriores, es mucho mayor que el tiempo requerido para crear un verdadero multidimensional y por lo tanto, el costo es mucho mayor.

Típicamente una aplicación OLAP llevará a cabo más procesos complejos que las aplicaciones relacionales normales. Un solo resultado, por ejemplo, “total de productos”, puede depender de un grado muy amplio de desagregación dentro de una base de datos. Por lo que contar con la arquitectura adecuada para cada aplicación es de suma importancia.

La innovación en el proceso de la toma de decisiones, se basa en facilitar un entorno mediante herramientas analíticas buen probadas y de técnicas que ayuden a determinar perspectivas claras sobre la dinámica del negocio [ARAB1997], [HAMM1996]. Las aplicaciones OLAP brindan la inteligencia de negocios requerida para tomar decisiones, y por consiguiente, definen la estructura y contenido del almacén de datos, el nivel histórico y de detalle requerido de la información. Finalmente, el proceso de recolección, filtrado e integración de los datos, obtenida de los OLTP, permite generar información valiosa propia del depósito de datos y los OLAP.

En la actualidad no podemos asegurar cuál estrategia o metodología es mejor, sin embargo, al analizar las tendencias generales del mercado, encontramos que la estrategia de desarrollar mercados de datos, está siendo adoptada más frecuentemente en los últimos tiempos

La necesidad de análisis multidimensional oportuno, como soporte para la toma de decisiones, es cada vez más creciente. Como respuesta los departamentos de sistemas se han inclinado por la tecnología de depósitos de datos para satisfacer las demandas de los usuarios. Sin embargo, es importante tomar en cuenta que existen diversas tecnologías orientadas a consultar, almacenar datos y analizar resultados.

En este sentido, los mercados de datos a base de OLAP, dentro o fuera de la arquitectura de depósitos de datos, ha demostrado ser una solución práctica para el usuario final.

Los conceptos introducidos en este capítulo conforman el conocimiento teórico necesario sobre los depósitos de datos, aspectos que son retomados en el capítulo tercero, donde se describe el procedimiento para la construcción de un depósito de datos.

Debido a que nuestro objetivo de investigación es desarrollar un procedimiento que permita la construcción de un depósito de datos que combine e integre aspectos que representen y manipulen información imprecisa por medio variable lingüísticas, se procede en el siguiente capítulo a plantear el problema del manejo de la información imprecisa e incompleta, a exponer las limitaciones de los sistemas tradicionales para su representación y a esquematizar nuestra propuesta para el manejo de este tipo de información.

CAPÍTULO 2

INFORMACIÓN INCOMPLETA E IMPRECISA

"Las oportunidades se disfrazan de trabajo duro, para que la mayoría de las personas no las reconozcan."

Ann Landers

La información "*incompleta*" responde a la necesidad de abordar situaciones reales que se presentan en el diseño, mantenimiento y explotación de las bases de datos. Con el término información "incompleta" se reúne todos aquellos casos en los que no se puede asumir un valor para un dato (valor ausente) o, no podemos asumirlo en forma precisa, sea cual fuere el motivo por el cual no poseemos dicho valor.

La información "*imprecisa*" forma parte de nuestra vida cotidiana y se manifiesta frecuentemente en cualquier acto de comunicación humana. Los aspectos más importantes de la información "imprecisa" que habitualmente manejamos son: la incertidumbre y la imprecisión en nuestras apreciaciones [KUYU1995]. El primero se deriva de apreciaciones realizadas sobre nuestra observación de la realidad que no pueden aportar un grado de certidumbre total sobre lo que afirmamos. El segundo aspecto se manifiesta a través del enunciado de conceptos que; no se encuentran bien diferenciados o definidos, o bien resultan subjetivos.

Estos enunciados proporcionan un grado de información sobre el mundo real, en algunos casos dicha información puede resultar suficiente y en otros puede no serlo, en cualquier caso aumentan nuestra información sobre el universo. A veces, los datos que caen dentro de una misma clase son tratados como iguales, indistintamente de su valor numérico individual. Con frecuencia las mediciones son de carácter cualitativo y están basadas en apreciaciones subjetivas del experto o tomador de decisiones.

Bajo estos conceptos, en el siguiente punto haremos énfasis en las propuestas de Codd y Date [CODD1970], [CODD1986], [CODD1990],[DATE1990] para mostrar las

limitaciones que presenta el modelo relacional para el tratamiento de la información incompleta o ausente.

En segunda instancia, introducimos los conceptos básicos de la lógica difusa [ZADE1965], [LEWO1989], [CHAR1994]. El planteamiento de la lógica y las matemáticas difusas será muy útil, no solo para capturar más semántica, sino para facilitar la manipulación de la incertidumbre por medio de las variables lingüísticas.

2.1 TRATAMIENTO DE LA INFORMACIÓN INCOMPLETA EN EL MODELO RELACIONAL

Históricamente las bases de datos han sido las herramientas diseñadas para llevar a cabo las tareas de almacenamiento y para proporcionar algunos de los mecanismos necesarios para el análisis de la información [GIRA1998].

El objetivo de una base de datos es el almacenar la información en forma conveniente, el permitir su modificación de forma segura y el de facilitar el proceso de recuperación de aquella información, que en un momento dado nos resulte necesaria, todo ello en un formato adecuado a nuestras necesidades [YAWO1998].

Como es bien sabido los estudios teóricos de bases de datos son la base fundamental para el desarrollo de buenos sistemas. La mayoría de estos estudios, están orientados al modelo relacional de bases de datos, cuya definición se fundamenta, en la lógica de primer orden.

Partiendo de la propuesta original descrita por Codd [CODD1970] para la definición del modelo relacional, se tiene que en principio un SABDR debe cumplir dos características básicas:

- Los datos son percibidos por el usuario como tablas o relaciones.

- Los operadores de los que dispone el usuario, por ejemplo para las consultas, generan como resultado de su aplicación nuevas tablas a partir de las existentes.

Asimismo, los conceptos más relevantes y los elementos más representativos de los SABDR tienen que ver con las siguientes partes bien diferenciadas: la estructura, integridad y manipulación de los datos. Para una descripción más amplia del modelo relacional, remitimos al lector al apéndice A.

El manejo de datos incompletos en una base de datos puede deberse a la pérdida de valores de algún atributo (al que se asigna entonces el valor desconocido), o a la ausencia del mismo en la vista que el sistema posee sobre los datos.

Codd [CODD1990] aborda el tratamiento de la información ausente desde la siguiente perspectiva semántica, empieza estableciendo una distinción entre que tipo de información se encuentra "ausente" y cual es el motivo por el cual se encuentra "ausente".

El primer aspecto trata de sí la información "ausente" es un registro, un atributo, o un conjunto de atributos. Dicha ausencia debe ser interpretada, por tanto, dentro de un contexto estructural.

El segundo aspecto, plantea el problema de sí la información no aparece porque se desconoce su valor en este momento, o porque dicho valor no es aplicable. Para modelar este aspecto, aboga por una representación basada en el empleo de dos tipos de "marcas" sobre los atributos susceptibles de admitir información ausente.

Para indicar que un valor se encuentra "ausente", en el sentido de que se desconoce su valor pero que es aplicable, emplea la marca de "valor ausente pero aplicable". Para representar que el motivo por el que se encuentra "ausente" un valor, es el que no es aplicable, propone el empleo de una marca que denote que la información se encuentra ausente y que además lo esta por que "no es aplicable".

El problema que plantea este tipo de tratamiento a la información conlleva una alteración sustancial a los aspectos básicos sobre los que se construye el modelo relacional. Dichos aspectos conciernen principalmente a la representación de este tipo de información en la estructura de una relación y sobre todo, a los aspectos relacionados con su manipulación. El último aspecto es el que representa un mayor problema, puesto que todas las propuestas que pretendan proporcionar mecanismos para manipular este tipo de información, se ven abocadas inevitablemente a modificar de forma sustancial los operadores algebraicos del modelo relacional.

Un ejemplo del primer tipo de información puede ser un atributo que recoja la fecha a partir de la cual las personas tienen trato con la organización, en una relación que almacena información sobre los mismos. Si para un cliente dado, desconocemos su fecha de inclusión, tenemos que señalar esta circunstancia mediante el empleo de la marca "valor ausente pero aplicable". El empleo del segundo tipo de marca se puede ilustrar mediante un ejemplo de atributo, que recoja información sobre la comisión por ventas que percibe cada empleado, si poseemos información en la base de datos sobre cual es el tipo de actividad que desarrolla cada empleado, esta claro que no se podrá proporcionar un valor para el atributo comisión por ventas, para un empleado que no se dedique a ventas, esta "ausencia" de información se denotara mediante la marca "valor ausente pero no aplicable" en el atributo comisión por ventas en la tupla correspondiente.

En el esquema de Codd la "ausencia" no es un valor del dominio considerado, si no un estado de la información, que se representa mediante el empleo de "marcas". Basa su esquema en una solución puntual que más bien parece ir dirigida a resolver los problemas de implementación, que a reconsiderar formalmente el modelo relacional para tratar de forma elegante el problema.

Por otra parte, C. J. Date presentó una aproximación para el tratamiento de los valores nulos [DATE1990], donde establece que el problema del tratamiento de valores nulos no se encuentra bien definido, y que por lo tanto, se debe incorporar esta

característica al modelo relacional, basada en el concepto que definió como “valores por defecto”.

En este sentido, han aparecido muchos autores que se han planteado la creación de sistemas mixtos en los que se aprovechen las posibilidades tanto del modelo relacional como de los entornos lógicos que permitan dar cabida a los diferentes tipos de información, esto es, valores desconocidos, valores no aplicables, valores imprecisos, así como derivar información adicional no almacenada explícitamente.

La posición de estos autores, es que el enfoque teórico más adecuado para tratar estos problemas, es la Teoría de Conjuntos Difusos, por lo que procederemos en el siguiente punto a tratar en más detalle estos conceptos.

2.2 INTRODUCCIÓN A LA LÓGICA DIFUSA

La Lógica, es la ciencia que estudia los principios formales del razonamiento. Según esta definición, la *lógica difusa* [ZADE1965] trata de los principios formales del razonamiento aproximado, en el cual, el razonamiento preciso, es un caso particular. La característica esencial de la lógica difusa, al contrario de lo que sucede en la *clásica*, es que permite modelar de alguna manera el razonamiento impreciso que, por otro lado, es la base del pensamiento humano. La idea básica es poder dar una respuesta aproximada a una pregunta, en función de unos hechos previamente almacenados y que pueden ser inexactos, incompletos o poco fiables.

Los grandes inconvenientes que plantea la lógica clásica para llevar a cabo este tipo de razonamiento son:

- En la *lógica bivaluada*, una proposición *p* es verdadera o falsa. En *lógica multivaluada*, una proposición puede ser verdadera, falsa, o tener un valor de verdad intermedio entre un conjunto finito de posibles valores de verdad. En

lógica difusa, se permite que los valores de verdad sean subconjuntos difusos definidos, generalmente, en el intervalo [0,1].

- En la lógica bivaluada, los predicados han de ser, necesariamente precisos, en el sentido de que no pueden describir subconjuntos difusos sobre el universo del discurso. En Lógica Difusa, los predicados pueden ser precisos (hombre, madre) o bien difusos (viejo, joven, bueno).

- En las lógica bivaluada y multivaluada, sólo se permiten dos cuantificadores: alguno y todos. En la lógica difusa, pueden utilizarse numerosos cuantificadores (la mayoría, pocos, muchos, normalmente, alrededor de, entre otros), que son vistos como números difusos que expresan la cardinalidad de algún conjunto difuso.

- La lógica difusa permite representar y manipular modificadores (difuso o no) de predicados, tales como: ligeramente, mucho, un poco, bastante, entre otros.

- En lógica bivaluada, una proposición puede ser cualificada asociándole un valor de verdadero o falso, mediante un operador modal como posible o necesario o bien mediante un operador intencional como creo, sé, o pienso. En lógica difusa, hay tres tipos de cualificación de predicados: *cualificación de la verdad*, por ejemplo "No es muy cierto que Alberto sea buen cliente", *cualificación de la probabilidad*, por ejemplo "Es poco probable que Alberto sea buen cliente" y *cualificación de la posibilidad*, por ejemplo "Es prácticamente imposible que Alberto sea buen cliente".

En resumen, la lógica clásica resulta restrictiva en cuanto a que:

- No ofrece los mecanismos adecuados para representar simbólicamente sentencias cuyo significado es impreciso.

- No ofrece un mecanismo de inferencia propicio para llevar a cabo el razonamiento aproximado.

L. Zadeh se basó en la lógica multivaluada de Lukasiewicz para formular su teoría de los conjuntos difusos y por ende de la lógica difusa [RASI1992]. La elección de L. Zadeh por la lógica de J. Lukasiewicz fue apoyada por los estudios de Maydole, que demostraron que la mayoría de las lógicas no estándares, daban lugar a paradojas en la teoría de conjuntos y de White que demostró que esto no ocurría con la lógica de J. Lukasiewicz [GAIN1983]. Así, da lugar a un cálculo de sistemas que abarca la imprecisión esencial de los datos del mundo real sin forzar la introducción de artefactos para satisfacer un requerimiento de precisión sin sentido.

La teoría de conjuntos difusos tiene lazos muy cercanos con las lógicas no estándar multivaluadas de J. Lukasiewicz, pero ha ido más allá de éstas en dos direcciones distintas:

- Al desarrollar una semántica lingüística para la lógica en términos de límites difusos.
- En su riqueza de aplicaciones prácticas, su real importancia es comparable sólo a la poseída previamente por el cálculo de predicados estándar

La lógica difusa, introducida por Lotfi Zadeh [ZADE1965] de la UC/Barkeley en los años sesenta, como un medio para modelar la incertidumbre del lenguaje natural es un superconjunto de la lógica booleana, que se ha extendido para manejar el concepto de verdad parcial (valores de verdad entre completamente verdadero y completamente falso).

Cae fuera del ámbito de este capítulo una descripción exhaustiva de la teoría de lógica difusa. Para más detalle remitimos al lector al apéndice B, en el cual se profundiza más sobre operaciones de conjuntos difusos y números difusos.

2.3 TRATAMIENTO DE LA INFORMACIÓN INCOMPLETA POR MEDIO DE VARIABLES LINGÜÍSTICAS

Así como hay una fuerte relación entre lógica booleana y el concepto de un subconjunto, hay una relación igualmente fuerte entre la lógica difusa y la teoría de conjuntos difusos [DUPR1992].

La interpretación original de conjunto difuso proviene de una generalización del concepto clásico de subconjunto ampliado a la descripción de nociones vagas e imprecisas [ZADE1965], [CHAR1994], [LEWO1989].

Esta generalización se fundamenta en que:

- La pertenencia de un elemento a un conjunto pasa a ser un concepto "difuso".
- Dicha pertenencia puede ser cuantificada por un grado, denominado habitualmente como "grado de pertenencia" de dicho elemento al conjunto y toma un valor en el intervalo [0,1].

Mediante estos conceptos, podemos representar de forma adecuada valores "imprecisos". En la tabla 2.1 se muestra la forma en que podemos modelar el concepto "joven". Para ello consideramos los valores que puede tomar la edad de un cliente (universo de discurso), y el grado de pertenencia de cada edad al concepto "joven".

Edad	Grado de pertenencia
20	1.00
25	1.00
26	0.96
28	0.74
30	0.50
35	0.20

Tabla 2.1 Conjunto difuso representando el concepto "joven"

Es necesario hacer notar que el concepto “joven” como otros muchos de naturaleza imprecisa responden a criterios subjetivos.

De forma más precisa podemos introducir la definición de conjunto difuso como sigue:

Definición 2.1. Un conjunto difuso A sobre un universo de discurso X es un conjunto de pares:

$$A = \{x, \mu_A(x) : x \in X, \mu_A(x) \in [0,1] \} \qquad (2.1)$$

donde $\mu_A(x)$ se denomina grado de pertenencia de x a A.

Según esto, si la “edad” es un universo de discurso de “joven”, el conjunto difuso que representa dicho concepto quedaría expresado en la forma:

$$\text{joven} = \{(20,1.00),\ldots(35,0.20)\}$$

El identificador “joven” con la connotación de que lleva asociado un conjunto difuso, recibe la denominación de “variable lingüística”.

Informalmente, una variable lingüística L, es una variable cuyos valores son palabras u oraciones en un lenguaje natural o en un subconjunto de él. Dada esta definición, se tiene los siguientes ejemplos (Tabla 2.2)

PERSONA	ALTURA(m)	Grado de ALTO
María	1,05	0,00
José	1,79	0,21
Juan	1,90	0,38

Tabla 2.2 Valor de pertenencia de un grupo de personas en el subconjunto difuso ALTO.

Expresiones como *a* es *x* pueden interpretarse como grados de verdad, por ejemplo, Juan es alto = 0.38.

Las funciones de pertenencia generalmente se presentan como si estuvieran basadas en un solo criterio, pero en la práctica eso no es tan cierto.

Una función de pertenencia para ALTO dependiente de la altura de la persona y su edad (es alta para su edad), es perfectamente legítima. Se refiere como una función de pertenencia bidimensional o una relación difusa. También, es posible contar con más criterios o que la función de pertenencia dependa de dos universos de discurso completamente diferentes.

Nótese que si se dan solo los valores 0 y 1 en estas definiciones, se obtiene las mismas tablas de verdad de la lógica booleana. Esto se conoce como el principio de extensión, el cual establece que los resultados clásicos de la lógica booleana se recobran de las operaciones de lógica difusa cuando todos los grados de pertenencia difusos se restringen al conjunto tradicional {0,1}. Esto establece a los conjuntos y lógica difusos como una generalización verdadera de la teoría de conjuntos y lógica clásicos. De hecho, a través de este razonamiento todos los conjuntos tradicionales son conjuntos difusos de un tipo muy especial; y no hay conflicto entre los métodos difusos y discretos.

Ejemplo 2.1. Suponga la definición anterior de ALTO y además disponga un subconjunto difuso VIEJO definido por la función de pertenencia:

$$\text{VIEJO}(X) = \begin{cases} 0 & \text{si EDAD}(x) < 18 \text{ años} \\ \dfrac{\text{EDAD}(x) - 18 \text{ años}}{42 \text{ años}} & \text{si } 18 \text{ años} \leq \text{EDAD}(x) \leq 60 \text{ años} \\ 1 & \text{si EDAD}(x) > 60 \text{ años} \end{cases}$$

Por brevedad, sea

a = x es ALTO AND x es VIEJO

b = x es ALTO OR x es VIEJO

c = NOT x es ALTO

Entonces se puede calcular los valores de pertenencia, según la tabla 2.3

PERSONA	**ALTURA (metros)**	**EDAD (años)**	**X es ALTO**	**X es VIEJO**	**a**	**b**	**c**
María	1,05	65	0,00	1,00	0,00	1,00	1,00
José	1,79	30	0,21	0,29	0,21	0,29	0,79
Juan	1,90	27	0,38	0,21	0,21	0,38	0,63

Tabla 2.3 Grados de pertenencia de personas en diferentes predicados difusos sobre altura y edad.

Para determinar los grados de pertenencia, los métodos se dividen en las siguientes categorías:

- *Evaluación subjetiva y sugerida*: las funciones de pertenencia las dan expertos en el área del problema, o se les dá un conjunto restringido de curvas de las cuales escoger. En métodos más complejos, los usuarios se analizan con métodos psicológicos.
- *Formas Ad Hoc*: son funciones de pertenencia creadas o empíricas.
- *Frecuencias o probabilidades convertidas*: algunas veces la información se toma en forma de histograma de frecuencia u otras curvas de probabilidad, que son usadas como la base para construir una función de pertenencia.
- *Mediciones físicas*: funciones de pertenencia cuyos datos de entrada son mediciones físicas. Se provee una función de pertenencia por otro método y los grados de pertenencia de los datos se calculan a partir de ellas.
- *Aprendizaje y Adaptación*: los expertos, usuarios o sistemas expertos ajustan las funciones de pertenencia a través de la retroalimentación.

Es muy común que en la información que manejan los humanos, existan calificativos vagos como: grande, importante, rápido, entre otros, que involucran incertidumbre, la cual se puede dar de modo explícito: casi cierto, muy posible, etc..., ó, de modo implícito, a través de expresiones condicionales: cuanto más joven es el paciente y cuanto más elevado es el índice de colesterol, más importante resulta pensar en un tratamiento serio. Es precisamente la gradualidad, la característica que imposibilita a la lógica de dos valores manipular adecuadamente este tipo de información [KYRU1995], [LEWO1989], [PAGR1995].

La información incierta y la información vaga se pueden agrupar bajo el término de *difusidad*, el cual se define como un tipo de imprecisión que surge al agrupar elementos en clases que no tienen fronteras bien demarcadas [ZADE1965].

Para ser más explícitos, el término difusidad abarca la posibilidad y la probabilidad. En la probabilidad se incluye información incierta, imprecisa e incompleta. En la posibilidad se incluye la información vaga, ambigua y ambivalente. La probabilidad puede manejar a la incertidumbre pero no a la posibilidad. Sin embargo, el enfoque posibilístico eventualmente, puede manipular la información incierta además de otros tipos de información difusa, debido a que la posibilidad da una estructura de orden, mientras que la probabilidad se limita a la propiedad de aditividad.

La información desconocida tiene tanto características probabilísticas como posibilísticas. Las siguientes descripciones de estos conceptos aclaran las diferencias entre ellos:

1. *Información Probabilística*.

    - *Incertidumbre*: se refiere a la ignorancia parcial de la especificidad de un cierto elemento de información y su descripción. Por ejemplo, la probabilidad de obtener un 6 al lanzar un dado no cargado es de 1/6. No hay certeza de que caerá un seis.

- *Imprecisión*: se refiere a la falta de especificidad de los contenidos de un elemento de información. Por ejemplo, una tasa de inflación entre 5 y 8 % anual.

- *Incompleta*: información que solo se conoce parcialmente.

2. *Información Posibilistica*

- *Vaguedad*: resulta de la falta de fronteras bien establecidas de un objeto, sea que se denoten por medio de números aproximados o por palabras. Por ejemplo, si se dice: Una tasa baja de inflación es buena, entonces la tasa de inflación es dependiente del contexto, pero indefinida aún en un solo contexto.

- *Ambigüedad*: información que no está claramente caracterizada y permite varias interpretaciones.

- *Ambivalencia*: caso particular de ambigüedad donde la información tiene, al mismo tiempo, dos interpretaciones posibles.

Frecuentemente, la difusidad se asocia con el concepto de vaguedad, sin embargo, la vaguedad es una característica dependiente de la decisión que se va a tomar siempre que la difusidad no.

Para manejar la probabilidad, la vaguedad, la ambigüedad, la incertidumbre y los datos incompletos, imprecisos o desconocidos se han propuesto varias herramientas: los métodos bayesianos, el modelo de factor de incertidumbre, la teoría de Dempster-Shafer, la lógica difusa, la teoría de conjuntos aproximados, entre otras.

Cada una de estas propuestas se destaca por manejar mejor alguno de los tipos de información anteriores y ser de utilidad limitada para los otros. Sin embargo, no existe

aún una teoría que pueda manipular por igual todos los tipos. En general, se considera que la herramienta más adecuada para manejar la vaguedad es la teoría de la lógica difusa y los conjuntos difusos aplicados a las bases de datos.

Dado que la lógica difusa es un formalismo matemático y un grado de pertenencia es un número preciso. La lógica difusa es una lógica de lo difuso, no una lógica que es en sí misma difusa. Así como las leyes de la probabilidad no son aleatorias, las leyes de la difusidad no son vagas [ZADE1965], [CHAR1994], [LEWO1989].

Comúnmente los valores difusos se mal interpretan como probabilidades, o la lógica difusa se interpreta como alguna forma nueva de manejar las probabilidades. Pero no es así. Un requisito mínimo de las probabilidades es la aditividad, esto es, deben sumar uno o el área bajo la curva de densidad debe ser uno.

En general, esto no es el caso con los grados de pertenencia y, mientras los grados de pertenencia pueden determinarse con densidades de probabilidad, también hay otros métodos que no tienen que ver con las frecuencias o las probabilidades.

Lo inverso si es cierto: toda distribución de probabilidad es un conjunto difuso. Puesto que los conjuntos y lógica difusos generalizan los conjuntos y lógica boléanos, ellos también generalizan la probabilidad.

En la probabilidad se supone que los eventos o las afirmaciones están bien definidos, lo que se cuestiona es la frecuencia con que esta afirmación va a ocurrir. Este tipo de incertidumbre o vaguedad se llama incertidumbre estocástica, opuesta a la vaguedad que concierne a la descripción del significado semántico de los eventos mismos, lo cual se llama difusidad.

Desde la perspectiva matemática los conjuntos difusos y la probabilidad existen como partes de una teoría de información mayor, la cual también incluye conjuntos aleatorios, teoría de evidencia Dempster-Shafer, intervalos de probabilidad, teoría de la

posibilidad, medidas difusas, etc. Aún más, se puede hablar acerca de eventos difusos aleatorios y de eventos aleatorios difusos. Este tema está más allá del alcance de este trabajo.

Finalmente, se pueden resumir las ventajas y desventajas de la lógica difusa como sigue:

Ventajas:

- No requiere constructores matemáticos complejos.
- Tiene buen fundamento matemático.
- Usa lenguaje natural (variables lingüísticas).
- Fácil de establecer.
- Arroja resultados exactos a partir de datos ambiguos.
- Captura mucha semántica.
- Trabaja bien en combinación con otras técnicas (algoritmos genéticos, redes de neuronas, sistemas expertos, programación orientada a objetos, etc...).
- Modela bien el pensamiento humano.
- Aplicable en múltiples disciplinas.
- Trata de ambigüedad y la vaguedad.

Desventajas:

- Se debe entender y ser capaz de definir el problema.
- No hay un método absolutamente fiable que defina la función de pertenencia.
- Se debe evaluar y afinar los resultados.
- Depende del contexto donde se ataque el problema.
- Existe cierta dependencia del experto.

En el próximo apartado explicamos la representación del conocimiento impreciso por medio de las variables lingüísticas, para que las mismas sean incorporadas en el

proceso de creación de un depósito de datos, con el fin de permitirle al tomador de decisiones analizar la información desde un punto de vista más cercano a los conceptos que él manipula, razón de nuestra propuesta de tesis.

2.4 REPRESENTACIÓN DEL CONOCIMIENTO IMPRECISO

Los diferentes elementos que forman parte del tratamiento impreciso pueden recibir diferentes representaciones. Así pues, por ejemplo, una distribución de posibilidad normalizada puede venir representada por diferentes tipos de funciones, nosotros emplearemos para la misma una representación trapezoidal, explicada en el apéndice B.

Pudiera parecer que el criterio de representación expuesto supone una fuerte restricción, sin embargo, esto no es así dada la naturaleza imprecisa de la información con la que tratamos. No parece muy acertado modelar unos datos, que en sí son imprecisos, con una representación extremadamente precisa dada por funciones no lineales.

A continuación describimos los criterios adoptados para la representación de datos imprecisos sobre referencial ordenado [ZADE1965]. Este grupo de datos contiene distribución de posibilidad sobre dominios continuos o discretos, sobre los que existe una relación de orden. Cada dato de este tipo tiene asociado una función de pertenencia dada por un experto.

Por simplicidad y eficiencia en el cálculo, adoptamos las siguientes representaciones para este tipo de datos:

- *Distribución de posibilidad trapezoidal normalizada*.

  Esta representación determina la función de pertenencia asociada al dato mediante el uso de cuatro parámetros: m, n, a, b. Ver figura 2.1.

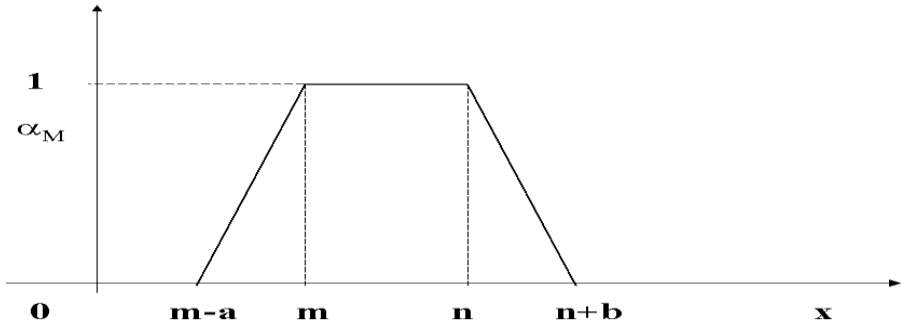


Figura 2.1 Distribución trapezoidal normalizada

- *Función de pertenencia.*

Si se habla de clientes y sus edades, se define un conjunto de clientes y un conjunto de EDADES, el cuál responderá a la pregunta ¿en qué grado el cliente x es joven, adulto o anciano?. Zadeh describe JOVEN como una variable lingüística, la cual representa nuestra categoría cognoscitiva de EDADES.

Para cada cliente se tiene que asignar un grado de pertenencia en el conjunto difuso EDADES, con la función de pertenencia basada en la edad del cliente, que se describe a continuación y el gráfico de esta función se muestra en la figura 2.2:

$$A_1(x) = \begin{cases} 1 & \text{si} \quad x \leq 20 \\ (35 - x)/15 & \text{si} \quad 20 < x < 35 \\ 0 & \text{si} \quad x \geq 35 \end{cases}$$

$$A_2(x) = \begin{cases} 0 & \text{si} \quad x \leq 20,\ x \geq 60 \\ (x - 20)/15 & \text{si} \quad 20 < x < 35 \\ (60 - x)/15 & \text{si} \quad 45 < x < 60 \\ 1 & \text{si} \quad 35 \leq x \leq 45 \end{cases}$$

$$A_3(x) = \begin{cases} 0 & \text{si} \quad x \leq 45 \\ (x - 45)/15 & \text{si} \quad 45 < x < 60 \\ 1 & \text{si} \quad x \geq 60 \end{cases}$$

- *Etiqueta lingüística.*

Los datos expresados mediante una etiqueta lingüística hacen referencia a un concepto impreciso, (a veces subjetivo), que lleva asociado una distribución de posibilidad. Por ejemplo, la etiqueta lingüística "edad", puede llevar asociada la distribución de posibilidad en representación trapezoidal que se muestra en la figura 2.2

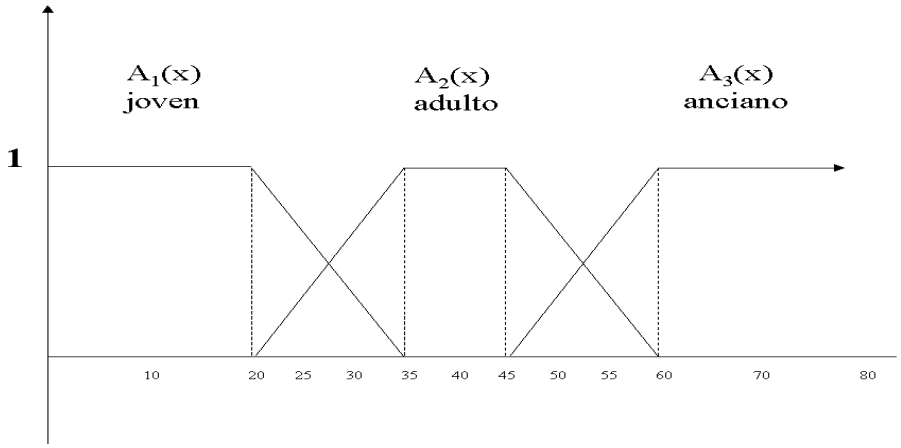


Figura 2.2 Etiqueta lingüística edad

El tratamiento de la información incompleta responde a la necesidad de abordar situaciones reales que se presentan en el diseño, mantenimiento y explotación de las bases de datos. Es por ello que definimos un modelo teórico con capacidad de representar información imprecisa lo suficientemente flexible y general como para permitir definir y utilizar reglas de carácter impreciso que serán aplicadas sobre los datos. Como propósito adicional y no menos importante se sientan las bases teóricas prácticas para la implementación de desarrollos específicos para el manejo de este tipo de datos.

Los conceptos expuestos, conforman el conocimiento teórico necesario para desarrollar el trabajo en los capítulos restantes. En el siguiente punto exponemos la implementación del conocimiento impreciso en una base de datos, para aquellos "datos imprecisos" que tienen etiquetas lingüísticas definidos sobre ellos.

2.5 MODELO PROPUESTO PARA EL MANEJO DE VARIABLES LINGÜÍSTICAS

Como hemos visto en el apartado anterior, existe cierto tipo de información sobre los atributos discretos que precisa ser almacenada de una forma accesible por el sistema. El MVL (Modelo para el manejo de Variables Lingüísticas) va a ser el encargado de organizar toda aquella información relacionada con la naturaleza imprecisa de estos atributos, el MVL será una extensión del catalogo del sistema, para ello organizaremos la información mediante el uso de tablas o relaciones.

Los elementos del tratamiento impreciso que se almacenan en el "MVL", son los siguientes:

- Que atributos de la base de datos reciben tratamiento impreciso.
- Que etiquetas lingüísticas hay definidos sobre cada atributo.
- Rangos de los valores difusos, y
- Descripción del grado de conocimiento del experto.

La organización de las entidades que constituyen el "MVL" se muestra en la figura 2.3.

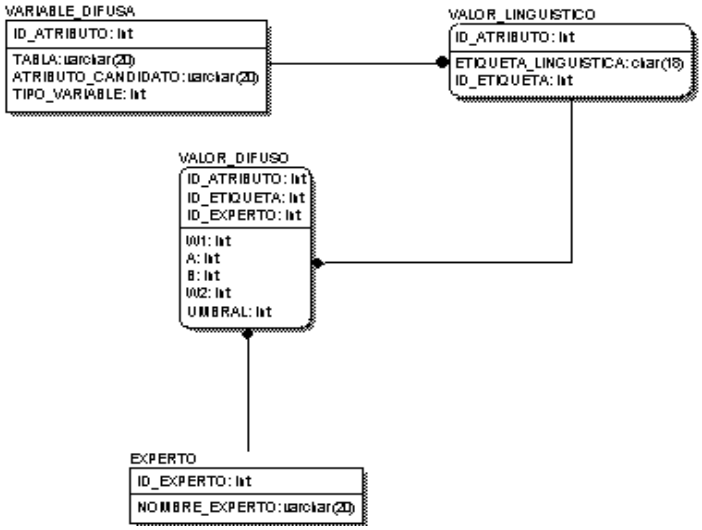


Figura 2.3 Modelo MVL

Procedemos a describir la estructura y significado de cada una de las entidades.

- *VARIABLE_DIFUSA*

Esta tabla contiene una descripción de aquellos atributos de la base de datos que son susceptibles de tratamiento difuso. Esta descripción se realiza en términos análogos a los empleados en los diccionarios de los RDBMS convencionales.

En primera instancia, la tabla consta de los siguientes atributos:

– ID_ATRIBUTO: Será de tipo numérico entero positivo y con un rango por determinar en función de las necesidades del sistema a implementar. Es la llave primaria de la tabla. Su misión será la de actuar en cualquier tabla de definición como referencia al campo al que va asociado, es decir, referenciamos a una columna difusa por su ID_ATRIBUTO.

– TABLA: Este campo es de tipo carácter y contiene el nombre de la tabla a la cual pertenece el campo difuso que se referencia en el atributo ATRIBUTO_CANDIDATO.

– ATRIBUTO_CANDIDATO: También es de tipo carácter. Los identificadores contenidos en este campo referencian a aquellas columnas que van a tener un tratamiento difuso, bien porque contengan información difusa, o bien porque, aún no teniéndola, pueden ser objeto de consultas difusas, la naturaleza de estos campos se ve reflejada en el campo TIPO_VARIABLE.

– TIPO_VARIABLE: Este campo tiene gran importancia, puesto que contiene información con respecto al tipo de datos y tratamientos que van a recibir las columnas referenciadas por el ID_ATRIBUTO. Inicialmente tomará el valor de “1”, indicando que se trata de una etiqueta lingüística.

- *VALOR_LINGÜÍSTICO*

Esta tabla contiene una lista de los objetos de tipo difuso que hay definidos en las columnas de la base de datos. La información se estructura como se detalla a continuación:

– ID_ATRIBUTO: Contiene el número que identifica a la columna sobre la que se define el objeto cuyo nombre aparece en el campo ETIQUETA_LINGÜÍSTICA de esta misma tabla. Constituye una clave externa a la tabla VALOR_DIFUSO.

– ETIQUETA_LINGÜÍSTICA: Contiene el nombre de la etiqueta lingüística asociada al atributo identificado por el valor ID_ATRIBUTO.

– ID_ITEQUETA: Asocia un número a cada objeto (etiqueta lingüística) que servirá para referenciarla en el resto de las tablas.

- *VALOR_DIFUSO*

Esta tabla contiene los puntos que determinan la función de pertenencia correspondiente a las etiquetas lingüísticas, como se muestra en las figuras 2.1 y 2.2. La información se estructura como se detalla a continuación:

– ID_ATRIBUTO, ID_ETIQUETA junto con el campo ID_EXPERTO, constituyen la llave primaria de esta tabla.

– W1: Numérico. $W1 = inf\{x: x \in soporte(ETIQUETA\_LINGÜÍSTICA)\}$. Ver apéndice B.

– A: Numérico. $A = inf\{x: x \in núcleo(ETIQUETA\_LINGÜÍSTICA)\}$. Ver apéndice B.

- B: Numérico. *B= sup{x: x ∈ núcleo(ETIQUETA_LINGÜÍSTICA)}*. Ver apéndice B.

- W2: Numérico. *W2= sup{x: x ∈ soporte(ETIQUETA_LINGÜÍSTICA)}.* Ver apéndice B.

- UMBRAL: Cuando se plantea una consulta en una base de datos con valores imprecisos se establecen una serie de condiciones a cumplir. Dada la naturaleza imprecisa de los operadores y de los datos sobre los que se opera, existe un grado de cumplimiento para cada condición involucrada en una consulta. Este grado de cumplimiento se halla comprendido entre cero y uno. Mediante el empleo de un umbral mínimo para el grado de cumplimiento podemos ejercer algún control sobre la precisión con que se satisfacen cada una de las condiciones de las consultas. Si establecemos un umbral de cumplimiento uno, para una condición envuelta en una consulta, eliminaremos aquellas tuplas o registros que no igualen o superen el umbral para esa condición. El umbral es el grado con el que se satisface el valor lingüístico; por ejemplo, si exigimos que el umbral con el que se satisface una condición simple sea 'VIEJO', estaremos indicando que aceptaremos valores con un grado igual o superior a 0.7. El valor del umbral es establecido en coordinación con los usuarios expertos, y tendrá al igual que las etiquetas lingüísticas, una connotación subjetiva.

- *EXPERTO*

Esta tabla contiene la codificación de los expertos utilizados en el modelo. Debido a que el valor del umbral que asociamos a cada valor lingüístico debe estar almacenado en el sistema, recurrimos a la evaluación subjetiva y sugerida de los expertos en el área del negocio:

- ID_EXPERTO: Número que identifica de forma única a un experto. Este campo forma parte la llave primaria en la tabla VALOR_DIFUSO.

– NOMBRE_EXPERTO: Nombre del experto que determina los grados de aceptación que se le va a dar a las variables lingüísticas.

Debido a que el objetivo principal de este trabajo es definir un modelo teórico lógico con capacidad para representar información imprecisa y lo suficientemente general como para permitir integrar las características más relevantes de cada una de las variables lingüísticas, en la siguiente subsección desarrollamos un ejemplo de la implementación de las variables lingüísticas a través de nuestro modelo MVL.

2.6 EJEMPLO DE IMPLEMENTACIÓN EN MVL

En este apartado vamos a ilustrar como se representan en la base de datos los elementos de la implementación introducidos en el punto anterior. Para ello vamos a emplear el ejemplo que se muestra en la tabla 2.4, que consiste en una relación que recoge datos sobre un conjunto de clientes. A lo largo de este ejemplo vamos a ver como se implementa la información que recoge dicha relación. Veremos la estructura interna que adoptan los campos difusos en la base de datos. Por último, mostramos como se actualiza la base del modelo MVL para albergar, la información relativa a los atributos y demás ítem difusos contemplados en dicha relación.

NOMBRE_CLIENTE	AGENCIA	EDAD	SERVICIOS
Juan	Heredia	30	2
Pedro	San José	60	1
Elena	Puntarenas	15	1
Julio	Limón	55	7

Tabla 2.4 Segmento de la tabla de clientes

Los atributos NOMBRE_CLIENTE y AGENCIA contienen información que no es sujeta a tratamiento difuso. Los atributos EDAD y SERVICIO los trataremos de forma difusa, y precisarán de la definición de los siguientes valores lingüísticos:

Nombre variable	Valor lingüístico
EDAD	Joven
	Adulto
	Anciano
SERVICIOS	Vinculación alta
	Vinculación media
	Vinculación baja

Una vez establecidos los valores lingüísticos asociados a los atributos sujetos a tratamiento difuso y determinado los rangos que permiten generar la función de pertenencia, procederemos a implementar las estructuras del modelo MVL, definidas en el punto 2.5, como se muestra en las tablas 2.5, 2.6 y 2.7.

ID_ATRIBUTO	TABLA	ATRIBUTO_CANDIDATO	TIPO_VARIABLE
01	Clientes	Edad	Numérica
02	Clientes	Servicios	Numérica
.....	.....	.....	.....
.....	.....	.....	.....

Tabla 2.5 VARIABLE_DIFUSA

ID_ATRIBUTO	ETIQUETA LINGÜÍSTICA	ID_ETIQUETA
01	Joven	01
01	Adulto	02
01	Anciano	03
02	Vinculación alta	01
02	Vinculación media	02
02	Vinculación baja	03
.....	.....	.....

Tabla 2.6 VALOR_LINGUISTICO

ID_ATRIBUTO	ID_ETIQUETA	ID_EXPERTO	W1	A	B	W2	UMBRAL
01	01	01	0	10	20	35	.8
01	02	01	20	35	45	60	.6
01	03	01	45	60	60	80	.7
01	01	02	0	10	20	35	.6
01	02	02	20	35	45	60	.7
01	03	02	45	60	60	80	.8
02	01	02	0	1	2	4	.6
02	02	02	3	5	7	10	.4
02	03	02	8	10	12	14	.2
...	...	...	..	..	..	..	..
..	...	...	..	..	..	..	..

Tabla 2.7 VALOR_DIFUSO

Basándonos en las estructuras anteriormente descritas, vamos a mostrar el proceso por medio del cual llegamos a la obtención de los resultados. Partiendo del esquema general para la generación de variables lingüísticas, mostrado en la figura 2.4, procedemos a describir mediante un ejemplo la resolución del algoritmo propuesto para la asignación de las etiquetas lingüísticas.

Como se ilustra en este esquema, se tiene dentro de las bases de datos de la organización, variables que son sujetas a tratamiento difuso, por ejemplo, la edad del cliente. Cada uno de estos atributos debe ser previamente catalogado en el MVL con base a los criterios brindados por los expertos. Estas definiciones son los recursos necesarios para que el algoritmo del MVL genere como resultado la correspondiente etiqueta lingüística asociada al valor original de entrada.

Seguidamente detallaremos los aspectos más relevantes de este algoritmo, para efectos de simplicidad, desarrollaremos el ejemplo para un único experto ID_EXPERTO = 01:

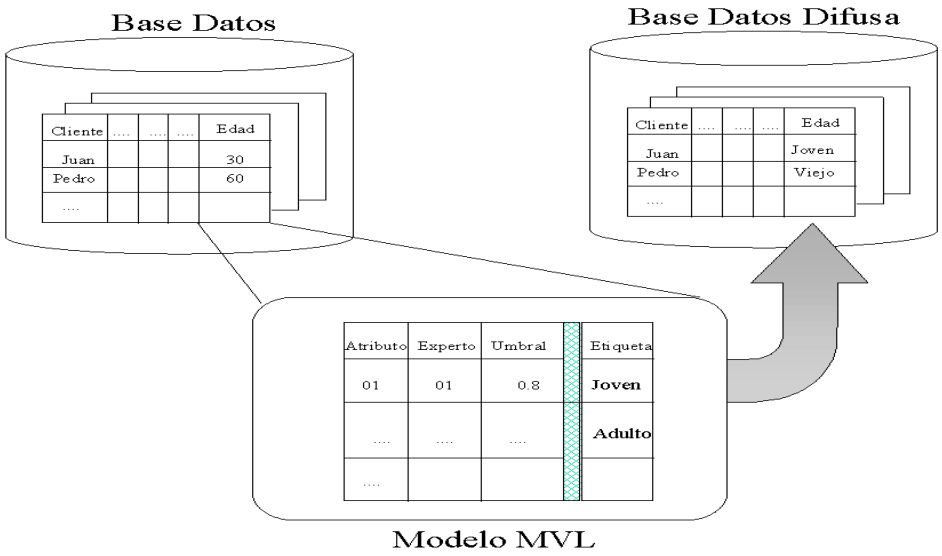


Figura 2.4 Esquema general para la generación de etiquetas lingüísticas

Algoritmo para determinar la etiqueta lingüística

1. Identificar en la entidad VARIABLE_DIFUSA, si el atributo es candidato a tratamiento difuso.

   Si es candidato: se obtiene el ID_ATRIBUTO, elemento necesario para determinar las variables lingüísticas y las funciones respectivas en el MVL.

   Ejemplo:
   ID_ATRIBUTO = 01 para la VARIABLE_DIFUSA (“edad”)

2. Con el ID_ATRIBUTO se recorre la tabla VALOR_LINGUISTICO, obteniendo los respectivos ID_ETIQUETA asociadas al atributo.

   Ejemplo:
   ID_ETIQUETA ={01, “joven”}, {02, “adulto”}, {03, “anciano”} para la VALOR_LINGUISTICO (id_atributo)

3. Para cada ID_ETIQUETA, ID_ATRIBUTO y ID_EXPERTO, se obtienen los valores w1, a, b, w2 y el umbral definidos en la tabla VALOR_DIFUSO.

   Ejemplo:

   W1 = 20 para VALOR_DIFUSO(id_atributo, id_etiqueta, id_experto)
   A = 35 para VALOR_DIFUSO(id_atributo, id_etiqueta, id_experto)
   B = 45 para VALOR_DIFUSO(id_atributo, id_etiqueta, id_experto)
   W2 = 60 para VALOR_DIFUSO(id_atributo, id_etiqueta, id_experto)
   UMBRAL = 0.6 para VALOR_DIFUSO(id_atributo, id_etiqueta, id_experto)

4. Evaluar la función de pertenencia basándose en el valor discreto asociado al ATRIBUTO_CANDIDATO y obtener el grado de pertenencia.

   Ejemplo:

   Valor pertenencia = 0.33 según función de pertenencia (w1, a, b, w2), para edad = 21

5. Comparar el valor de pertenencia con su respectivo umbral, con la finalidad de seleccionar la ETIQUETA_LINGUISTICA que será asociada al ATRIBUTO_CANDIDATO.

   Ejemplo:

   Edad = 21 el valor de pertenencia es 0.9

   Si el valor de pertenencia (0.9) es mayor al UMBRAL (0.8) se toma la ETIQUETA_LINGÜÍSTICA de la tabla VALOR_LINGUISTICO (ID_ETIQUETA, ID_ATRIBUTO), en este caso la etiqueta asociada al atributo candidato será JOVEN.

6. Si no se ha obtenido la ETIQUETA_LINGÜÍSTICA, se debe obtener el siguiente ID_ETIQUETA, y proceder con el paso 5 hasta obtener la correspondiente etiqueta lingüística.

En las tablas 2.8 y 2.9, se muestran los diferentes valores obtenidos al aplicar el algoritmo para umbrales definidos por diferentes expertos.

Edad	A1(X)	A2(X)	A3(X)	Etiqueta
	Umbral: 0.8	Umbral: 0.6	Umbral: 0.7	
15	1	-	-	Joven
21	0.9	-	-	Joven
25	0.66	0.33	0	Adulto
27	.53	0.46	0	Adulto
30	0.33	0.66	-	Adulto
31	0.26	0.7	-	Adulto
50	0	0.66	-	Adulto
54	0	0.4	0.6	Adulto
57	0	0.2	0.8	Anciano
60	0	0	1	Anciano

Tabla 2.8 Información del experto 1

Edad	A1(X)	A2(X)	A3(X)	Etiqueta
	Umbral: 0.9	Umbral: 0.8	Umbral: 0.7	
15	1	-	-	Joven
21	0.9	0.06	0	Adulto
25	0.66	0.33	0	Adulto
27	0.53	0.46	0	Adulto
30	0.33	0.66	0	Adulto
31	0.26	0.7	0	Adulto
47	0	0.8	0.13	Adulto
50	0	0.66	0.33	Adulto
54	0	0.4	0.6	Adulto
57	0	0.2	0.8	Anciano
60	0	0	1	Anciano

Tabla 2.9 Información del experto 2

Como puede notarse para una misma edad, por ejemplo, 21 años, la función le asigna, la etiqueta de JOVEN para el experto 1, con un umbral de aceptación de 0.8, y la etiqueta ADULTO para el experto 2, con un umbral de aceptación de 0.9.

Por lo tanto podemos concluir que el procedimiento establecido en nuestro modelo MVL nos permitirá incorporar en un depósito de datos información incierta, la cual esta presente en muchas áreas de la actividad humana.

A pesar de que se han propuesto varias herramientas: los métodos bayesianos [PLAN1991], el modelo de factor de certidumbre [VAN1990], la teoría de Dempster-Shafer [PLAN1991], la teoría de conjuntos aproximados [PASL1986] y otros, consideramos que la utilización de las variables lingüísticas [ZADE1965] modelándolas con la noción de grado de membresía, permiten representar adecuadamente la incertidumbre, haciendo que el tomador de decisiones (experto) visualice los datos de forma más natural, dando como resultado de está investigación una innovación tecnológica importante.

En el siguiente capítulo se procede a detallar las actividades y elementos involucrados en la construcción de un depósito de datos. Para que al producto final, se le aplique el modelo MVL, permitiendo de esta forma generar un depósito de datos con características difusas, objetivo principal de nuestra investigación. La aplicación y unificación de estos conceptos son desarrollados mediante un caso práctico en el capítulo cuarto.

CAPÍTULO 3

METODOLOGÍA PARA LA CONSTRUCCIÓN DE UN DEPÓSITO DE DATOS

"Una arquitectura, no puede por ella misma mejorar la calidad o flexibilidad de los sistemas operacionales existentes ni tampoco puede resolver ninguno de los problemas de la calidad de los datos."

Tom Hammergren

Con el surgimiento de los depósitos de datos las organizaciones necesitan procedimientos que les permitan administrar de forma practica la transición a estas nuevas tecnologías, la estructura que organiza todo este proceso será llamada arquitectura. Este proceso es particularmente complejo cuando se involucran tareas que buscan la integración de las aplicaciones operacionales con las nuevas tecnologías de depósitos de datos.

El termino arquitectura tiene muchos significados para las personas que desarrollan y utilizan sistemas computacionales. Literalmente podemos definir la arquitectura como un estilo y un método para diseñar y construir un arreglo de partes ordenadas [HAMM1996].

La *arquitectura* debe proveer el conjunto de lineamientos, necesarios para estructurar el plan general de construcción del depósito de datos. Abarcando áreas específicas como son las labores de planeación, administración y control, hasta la etapa de diseño e implementación. Además, debe proveer un diseño que explique como la visión, las metas, y los objetivos del depósito de datos serán implementados [HAMM1996], [GIRA1998].

Es por esta razón que en este capítulo desarrollamos una arquitectura que sirva de metodología en el proceso de construcción de un depósito de datos, objetivo principal de nuestro trabajo de investigación. La figura 3.1 muestra de forma general las áreas abarcadas por nuestra metodología.

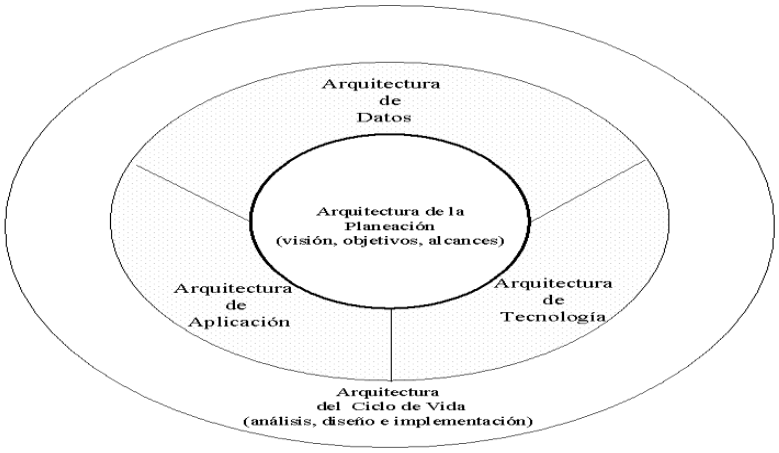


Figura 3.1 Áreas de la arquitectura

El proceso para elaborar esta arquitectura involucra las siguientes fases:

3. *Arquitectura de la Planeación:* fase en la que se define el problema así como los alcances, objetivos y beneficios del proyecto. Es por ello importante la formación de un grupo interdisciplinario con amplios conocimientos del negocio y de la plataforma tecnológica de la organización, para que en conjunto estructuren el plan general del proyecto.

4. *Arquitectura actual:* etapa en la que se examina el estado actual de la plataforma tecnológica de los sistemas que estén dentro del alcance del proyecto, identificando la arquitectura de las aplicaciones y datos, así como sus relaciones.

5. *Arquitectura del ciclo de vida:* fase final de nuestra metodología, que involucra las tareas relacionadas con las labores de diseño, construcción e implementación del depósito de datos

La base de nuestra propuesta está formulada a través de muchas revisiones de otras arquitecturas, incluyendo aquellas que provienen de organizaciones con una orientación tecnológica, por ejemplo: IBM, DEC, NCR, ORACLE y MICROSOFT.

En los siguientes apartados se procede a detallar cada una de estos elementos. En el apéndice F, se detalla una guía de la metodología.

3.1 ARQUITECTURA DE LA PLANEACIÓN

El elemento más importante en la construcción de un depósito de datos es entender claramente el tipo de decisiones que se requieren hacer y la información requerida para soportar éstas decisiones, razón por la cual es necesario una definición adecuada de la visión, objetivos y alcances del proyecto.

Debido a que los depósitos de datos están constituidos para soportar la toma de decisiones del negocio y no para automatizar procesos específicos de la organización, estos conceptos deben ser orientados en función de los requerimientos del usuario y tener un enfoque dirigido al negocio [BOAR1996], [GIRA1998], [HAMM1996].

El equipo de trabajo involucrado en la definición de estos aspectos debe estar conformado por especialistas en tecnología y expertos del negocio, los cuales deben tener claro los siguientes puntos [GIRA1998], [HAMM1996]:

- La arquitectura de la tecnología de información existente.
- Entender los problemas y necesidades de información de los niveles superiores, especialmente el modelo de datos y las prioridades del negocio.
- Organizar la información operativa de las áreas involucradas y los sistemas que mantienen estos datos.

Esto dará como resultado un mejor entendimiento de los tipos de decisiones que se requieren y de la información necesaria para el soporte a la toma de decisiones.

3.1.1 Crear una visión

A finales de los 80 las organizaciones de manufactura comenzaron a implementar sus nuevas visiones con conceptos de puntualidad y eficacia [HAMM1996], la visión de puntualidad cambia dramáticamente los modelos empresariales. Bajo este esquema los proveedores pueden ayudar a las empresas, a vender sus productos reabasteciendo los inventarios mientras lo requieran los distribuidores. Está visión ayuda en el desarrollo de los nuevos sistemas y al personal de la empresa a concentrarse en mejorar el proceso del negocio [HAMM1996].

Cuando las compañías crecen y llegan a convertirse en líderes dentro de su propia disciplina los procesos del negocio y la estructura organizativa comienza a evolucionar, esta transformación conduce a que las organizaciones replanteen sus sistemas de apoyo a decisiones [GIRA1998,HAMM1996].

En el contexto de depósitos de datos la visión debe ser enfocada en apoyar el proceso de tomas de decisiones más que a automatizar los procesos específicos del negocio [HAMM1996]. Este tipo de visión produce un sistema de información que ayuda al usuario a ejecutar más eficientemente sus tareas: los sistemas empiezan a ser parte de sus procesos más que a inhibir sus labores.

3.1.2 Definir alcances del proyecto

Dado que la idea principal de un depósito de datos es permitir a la organización compartir datos, el termino “organización” debe incluir todas las áreas que necesiten compartir cantidades sustanciales de información. Este compartimiento de datos, debe comprender el acceso constante y el uso del depósito de datos y no simplemente la consolidación periódica de los datos [HAMM1996], [YAWO1998].

Cuando el alcance es muy limitado, tal como un simple departamento, se tiene como resultado una arquitectura incompleta y carente de fortalezas para soportar otras áreas de la empresa. De otra forma cuando el alcance es demasiado extenso, tal como la empresa entera, se pone al equipo de trabajo en serios aprietos, en los cuales no se va a tener el suficiente tiempo y recurso para obtener una definición suficientemente detallada de la arquitectura, lo que la hace inválida.

Un buen alcance debe incluir todas las funciones del negocio como también las responsabilidades enmarcadas dentro de la frontera del depósito de datos. Con este tipo de alcance los beneficios económicos y la justificación del desarrollo de una arquitectura pueden ser justificados [BAUM1996], [HAMM1996].

3.1.3 Definir metas y objetivos del proyecto

Dentro de esta sección, el plan de la arquitectura debe establecer claramente la razón de la iniciativa del depósito de datos. Debe quedar definido el uso que se le dará al depósito de datos a través de la empresa como también los beneficios que el negocio obtendrá de tal aplicación [HAMM1996].

Es importante definir estas metas y objetivos de forma simple, concisa y un lenguaje no técnico. Las metas y objetivos definen los logros específicos deseados del esfuerzo de desarrollo [HAMM1996], [YOUN1994], [YAWO1998].

3.1.4 Estructurar el grupo de trabajo

Nuestra metodología provee al equipo de trabajo la estructura general de la arquitectura, sin embargo, es importante definir la composición del grupo de trabajo, los roles y responsabilidades de cada uno de los miembros.

El equipo de trabajo debe incluir al menos los siguientes puestos y responsabilidades [GIRA1998], [HAMM1996], [INGL1997]:

- *Administrador del proyecto:* administra las actividades diarias del proyecto y de los miembros del grupo, además elabora la información pertinente para evaluar el estado del proyecto.

- *Analistas de negocios:* provee desde el punto de vista del usuario el conocimiento del negocio.

- *Arquitectos:* este grupo debe estar conformado por especialistas en la parte de datos, aplicación y tecnológica. Cada uno de ellos provee con el conocimiento de su área de especialización.

- *Administradores del conjunto de herramientas:* la tecnología de depósitos de datos esta conformada por un conjunto de herramientas, por lo que se requiere de una persona que se encargue de las labores de instalación, preparación y el uso de estos productos.

En el proceso de desarrollo de software, las personas son el ingrediente más importante. El objetivo es formar el mejor grupo de trabajo para el proyecto de desarrollo del depósito de datos, para ello una habilidad importante a considerar es el conocimiento que tengan las personas de las áreas del negocio contempladas dentro del alcance del proyecto.

Cuando se inician las tareas de construcción de un primer depósito de datos, el proceso de desarrollo es inmaduro, razón por la cual se debe trabajar con un equipo base, que adquirirá experiencia y madurez conforme avance el proyecto. Este conocimiento podrá ser aprovechado por la organización en el desarrollo de nuevos proyectos.

3.1.5 Plan de Trabajo

El grupo de trabajo, guiado por el administrador del proyecto, debe elaborar el plan de trabajo para todas las actividades involucradas en el proceso de construcción del depósito de datos [HAMM1996], [INGL1997]. En este plan se define el estimado de tiempo y los recursos requeridos para realizar cada una de las tareas, además de los puntos de control del proyecto.

La arquitectura de planeación permite establecer las reglas que regirán el proceso de construcción del depósito de datos. Sin embargo, no puede por ella misma mejorar la calidad o flexibilidad de los sistemas existentes, ni tampoco puede resolver ninguno de los problemas de la calidad de los datos en las bases de datos [HAMM1996].

Algunos de los beneficios que se obtienen al utilizar este esquema son:

- *Datos más consistentes:* como es común los estándares de datos y modelos se vuelven ampliamente usados, la consistencia de los datos se volverá la norma dentro de su empresa. Una buena arquitectura de datos emplea estándares de datos consistentes y modelos que están completamente catalogados.

- *Simplifica el desarrollo de aplicaciones:* el diseño inteligente y la implementación con base a un enfoque cohesivo de cómo la empresa crea, accede y modifica los datos, permite que el tiempo de implementar nuevas aplicaciones o cambiar las existentes se vea reducido.

- *Grado de reacción del negocio:* la integración de los datos, las aplicaciones y la arquitectura tecnológica permitirá a la empresa responder a las necesidades del negocio de forma rápida, consistente y de alta calidad.

Una adecuada planeación ayudará a aclarar los requerimientos del depósito de datos. Es por ello que esta arquitectura debe ser establecida y aceptada anticipadamente por la organización antes de iniciar con su desarrollo [HAMM1996].

Como siguiente actividad en el proceso de construcción de un depósito de datos, en el siguiente apartado se desarrollará la arquitectura actual, en la cual se modela los sistemas actuales de la organización y sus tecnologías

3.2 ARQUITECTURA ACTUAL

Muchas metodologías proveen procesos que permiten definir, construir e integrar una arquitectura que soporte el esfuerzo del desarrollo de aplicaciones, incluyendo datos y tecnología [GIRA1998], [HAMM1996], [INGL1997]. Estas metodologías incluyen todas las tareas y actividades que deben ser ejecutadas, como también la información que debe ser capturada para el éxito e implementación de la nueva arquitectura.

Es por ello, que se debe seleccionar una metodología que permita un entendimiento más amplio de la arquitectura actual de la empresa, capturando de cada sistema al menos la siguiente información [HAMM1996]:

- Nombre del encargado del sistema.
- Función del negocio o departamento soportado por la aplicación.
- Una definición de cómo el sistema ayuda a la función empresarial.
- Etapa del ciclo de vida del sistema: planeación, desarrollo e implementación.
- Departamento y número de personas responsables del mantenimiento de la aplicación.
- Tecnología utilizada por el sistema: plataforma de hardware, comunicación y software.
- Documentación relacionada con el sistema: diagramas, modelos y flujos.

Para efectos de nuestra investigación, el análisis de la arquitectura de la organización contempla los siguientes aspectos:

- *Arquitectura de aplicación:* define y soporta los procesos de software para la implementación de los requerimientos funcionales del negocio.

- *Arquitectura de datos:* consiste en la organización de las fuentes de información y en el almacenamiento de los datos del negocio a lo largo de la empresa.

- *Arquitectura de tecnología:* es la infraestructura conceptual que permite que los datos y aplicaciones interactúen apropiadamente a lo largo de toda la empresa.

La metodología debe permitir integrar estos tres elementos en un único ambiente, por qué la interacción de éstos son la llave para implementar una arquitectura empresarial que soporte el dinamismo del negocio. A continuación procederemos a ilustrar cada uno de estos conceptos.

3.2.1 Arquitectura de aplicación

Cuando se evalúa la arquitectura actual de la empresa, se debe inventariar cada función del negocio requerida dentro del alcance del depósito de datos. Estas funciones deben ser modeladas y relacionadas con cada uno de los sistemas, para obtener un modelo donde se define las relaciones e interdependencias entre cada uno de ellos [HAMM1996], [INGL1997].

Se puede determinar rápidamente cuales son las fuentes de datos necesarias a ser analizadas basándonos en el enfoque de aplicaciones. Este enfoque captura la información general del sistema como también las relaciones con otras aplicaciones, como se muestra en la figura 3.2. Mapa de los principales sistemas y sus relaciones.

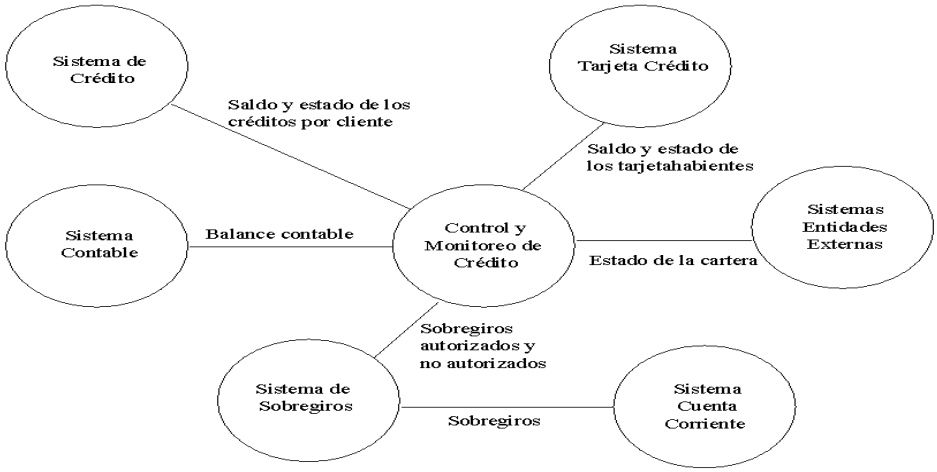


Figura 3.2 Mapa de los principales sistemas y sus relaciones

Además, la información obtenida debe permitir construir un índice comprensivo de las aplicaciones, ver tabla 3.1 índice de aplicaciones, donde se identifican los sistemas y las funciones del negocio que ellos soportan. Estos datos empiezan a ser relevantes para el equipo de desarrollo en la definición del tema.

Nombre del sistema	Función del negocio
Crédito	Formalización y asignación del crédito Gestión de cobro
Tarjeta Crédito	Afiliación y venta de Tarjeta Control y administración
Contable	Balance general Auxiliar contable
Sobregiros	Administración de sobregiros Gestión de cobro
Cuenta Corriente	Apertura de Cuentas Administración de la Cuenta Corriente Intereses
Entidades Externas	Estado de la Cartera

Tabla 3.1 Índice de aplicaciones

3.2.2 Arquitectura de datos

Un almacén de datos es una aplicación para el manejo de datos [INMO1995], por lo tanto es importante contar con un modelo de datos que muestre cada una de las áreas

del negocio. El modelo de datos debe enfocarse en la información más importante y la relación entre cada una de las entidades [HAMM1996].

Los diagramas de entidad - relación (ERD) son los gráficos más frecuentemente asociados con el modelamiento de datos. Es una representación conceptual de los objetos del mundo real y las relaciones entre ellos, típicamente se define la información de un sistema mediante el agrupamiento de los ítem en entidades lógicas [HAMM1996].

Las entidades representan las principales áreas de una aplicación y los procesos del negocio. La relación entre las entidades provee detalle acerca de cómo los datos fluyen. Para mayor información de este modelo ver [CODD1970], [CODD1990], [CHEN1976]. En la figura 3.3 se ilustra un ejemplo de datos a un nivel muy general utilizando el modelo entidad – relación.

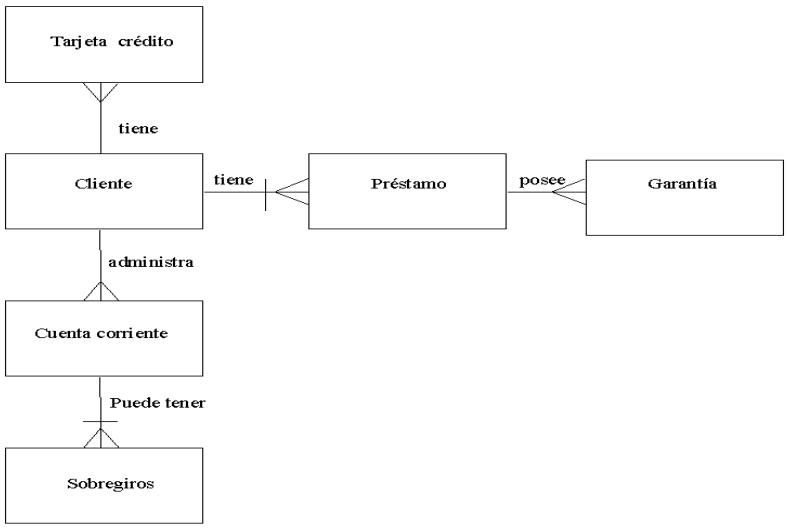


Figura 3.3 Ejemplo de un diagrama entidad-relación

La relación se nombra típicamente con verbos o acciones que permiten al lector del diagrama (entidad- relación) visualizar una acción, tales como:

- Un cliente “tiene” uno o muchos prestamos.
- Un cliente “tiene” una o varias cuentas corrientes.

- Una cuenta corriente “puede tener” un sobregiro.
- Un préstamo “posee” garantías.

Cuando se desarrolla éste diagrama se abstrae al nivel más alto las entidades, por ejemplo la entidad préstamo se puede descomponer en una serie de entidades en las que se muestre el detalle de los pagos realizados, plan de pago y otros. Por lo que es importante que el modelo no pierda consistencia al bajar de nivel.

3.2.3 Arquitectura de la tecnología

La arquitectura tecnológica define las principales tecnologías y plataformas que proveen un ambiente para las aplicaciones que administran datos [HAMM1996]. La plataforma tecnológica provee el medio para recolectar los datos de las diferentes fuentes de información. Además de almacenar, procesar, transportar y entregar los datos a los clientes.

Durante esta fase, se debe definir todas las plataformas que actualmente soportan las aplicaciones y datos previamente modelados. Este modelo es similar en naturaleza al modelo de datos y aplicaciones. Este enfoque incluye la identificación de los tipos de clientes y servidores utilizados por los usuarios de los sistemas de información.

El modelo debe incluir una definición del esquema de conectividad utilizado por la empresa, esto es, la estrategia de red e interconectividad entre los sistemas. La figura 3.4 ilustra de manera general una arquitectura tecnológica.

Es importante que se detalle toda la información que sea relevante, para establecer apropiadamente las bases a partir de las cuales se va a construir el depósito de datos. Esto incluye comprender mejor el software que maneja los recursos de datos de la Institución, así como los métodos de interconectividad entre ellos, lo que conlleva a una evaluación más exhaustiva de los principales sistemas de aplicación.

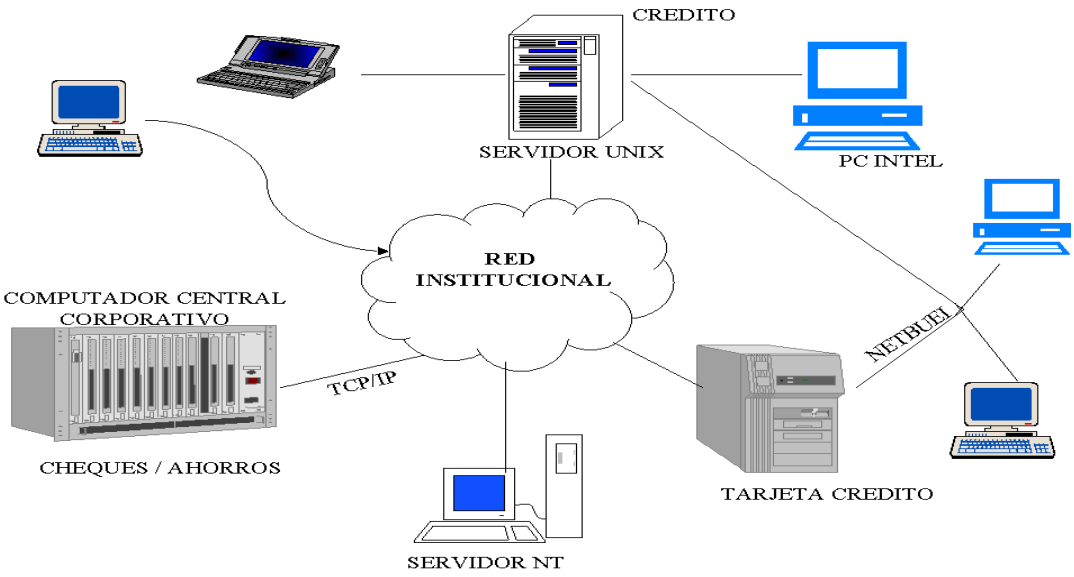


Figura 3.4. Arquitectura general de tecnología

La tabla 3.2 muestra la documentación de las aplicaciones y sus respectivos sistemas de administración de base de datos. Cada uno de los sistemas de aplicación es anotado en la primera columna y en las siguientes columnas se identifica el respectivo administrador de base de datos asociado a cada aplicación.

Aplicación	ORACLE	DMS II	OTROS
Crédito	✔		
Cuenta corriente		✔	
Tarjeta crédito	✔		
Sobregiros		✔	
Contabilidad	✔		
Entidades externas			✔ (Excel)

Tabla 3.2 Administradores de bases de datos

3.3 ARQUITECTURA DEL CICLO DE VIDA

Las principales fases a desarrollar en el proceso de construcción de un depósito de datos [GIRA1998], [HAMM1996], [KIMB1998] se ilustran en la figura 3.5.

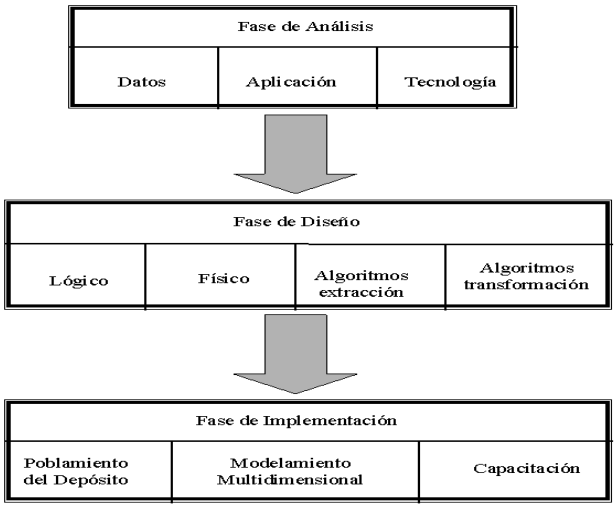


Figura 3.5 Arquitectura del ciclo de vida del desarrollo de un depósito de datos

Estas fases se centran en definir los detalles del alcance del proyecto. En ellas el equipo de trabajo empieza a desarrollar respuestas a:

- ¿Qué información se contendrá dentro del almacén de datos?.
- ¿Qué áreas alternativas de información son posibles entregar al usuario?.
- ¿Cómo serán identificados los riesgos y como se resolverán para minimizar el impacto en el proyecto?.
- ¿Quién será el sector meta y que detalles se conocen de él?.

El conocimiento de esta información proporcionará al grupo de trabajo la inteligencia necesaria en las áreas contempladas en el alcance del proyecto. Su análisis debe enfocarse en definir [HAMM1996]:

- ¿Qué datos son requeridos por los usuarios?.
- ¿Qué datos serán de apoyo para la toma de decisiones?
- ¿Qué grado de conocimiento requiere el equipo de desarrollo para generar la información al usuario final?.
- ¿Cuál es el plan de actualización y mantenimiento de los datos?

- ¿Se puede obtener la información requerida por los usuarios a partir de las fuentes de datos de la organización?.

3.3.1 Fase de análisis

Esta fase inicia con un claro entendimiento de las necesidades de las áreas funcionales de la organización, lo cual asegura que el producto obtenido cumple con los requerimientos y expectativas de la organización.

Los siguientes son ejemplos de requerimientos de un depósito de datos:

- Datos de múltiples áreas de la organización deben ser integrados con el objetivo de facilitar una visión única de la empresa a los usuarios.
- La empresa debe responder rápidamente a cambios en el mercado, incluyendo nuevos productos, nuevos competidores y nuevos clientes.
- La información debe asistir a las unidades del negocio en ejecutar más efectivamente sus tareas de mercadeo y venta.

Cada uno de estos requerimientos provee detalles importantes que ayudan en la definición de la tecnología que puede ser utilizada en la implementación final de la solución. El realizar este análisis conlleva a que el equipo de desarrollo tenga una visión clara de las necesidades tecnológicas requeridas para el depósito de datos [HAMM1996].

Dentro de este contexto, el grupo de trabajo debe abocarse en analizar los requerimientos desde los siguientes puntos de vista [GIRA1998], [HAMM1996]:

Área de análisis	Detalle de requerimientos
Datos	Optimización de consultas
	Tamaño de los datos
	Integración de múltiples formatos

Área de análisis	Detalle de requerimientos
	Calidad de los datos
Aplicación	Herramientas de diseño
	Interfases críticas
	Herramientas de prueba
	Herramientas de usuario final
Tecnología	Centralizado o distribuido
	Servicios de red
	Seguridad y autenticación
	Procesos de cliente – servidor
	Almacenamiento en línea y fuera de línea

Esta fase es crítica para el éxito del depósito de datos, en ella los requerimientos deben quedar bien definidos en termino de:

- *Área y tema de investigación:* la selección cuidadosa de estos elementos permitirá identificar las distintas unidades de interés. Por ejemplo: Área de Crédito, Mercadeo, Ventas, entre otros.

- *Granularidad:* se refiere al nivel de detalle de la información requerida, a menor Granularidad, mayor cantidad de detalle. Para incrementar su granularidad los datos operacionales deben resumirse y acumularse todavía más, por lo general entre mayor sea la granularidad mayor será la cantidad de procesamiento requerido para convertir y resumir los datos.

- *Dimensiones:* un depósito de datos permite organizar los datos operacionales en múltiples dimensiones. Las más importantes son; el tiempo, geografía, cliente, producto y organización. Por ejemplo, la dimensión de tiempo conlleva la fijación de la fecha y hora de la información, la dimensión de geografía define el país, región y distrito. Con la dimensión de producto se

deben especificar la familia y línea de los productos. La dimensión de organización estructura la información desde el punto de vista de la organización, departamentos, unidades y secciones.

El grupo de trabajo debe trasladar los requerimientos en metas que correspondan a una de las siguientes áreas: datos, aplicación y tecnología.

Datos

La primera labor a realizar en el análisis de los requerimientos es la identificación de cada una de las fuentes de datos. Estas fuentes pueden ser sistemas operacionales, mercados de datos departamentales o base de datos multidimensionales. Cada una de ellas debe ser analizada y documentada separadamente con el objetivo de facilitar la implementación física de los requerimientos [GIRA1998], [HAMM1996].

Esta labor de documentación se realiza por medio de los metadatos. Por lo general, los metadatos se definen como datos acerca de los datos [BAUM1996], [GIRA1998], [INMO1995], [YAWO1998], los metadatos es la representación de los diversos objetos que definen una base de datos. En una base de datos relacional, esta representación consistiría en las definiciones de tablas, columnas, tipo de datos y otros aspectos generales.

En nuestro caso usaremos el termino de metadato para hacer referencia a todo lo que defina un objeto en el depósito de datos, ya sea una tabla, una columna, una regla de negocios o una transformación. La comprensión de estas definiciones es esencial para todos los aspectos del desarrollo del depósito de datos, desde el desarrollo de programas de extracción de las bases de datos fuentes que alimentan al depósito de datos, hasta la transformación de datos de múltiples bases de datos, para que puedan almacenarse en un formato común dentro del depósito de datos.

Los estándares definidos hasta ahora comprenden el Estándar de Diccionario de Información de Recursos ANSI (IRDS), también se ha establecido un Consejo de

Metadatos para definir un estándar mínimo que permite el intercambio de metadatos entre los productos de administración de metadatos que ofrecen diversos fabricantes [GIRA1998]. Para efectos de nuestra investigación, la tabla 3.3 ilustra el esquema empleado para la documentación del metadato.

Información de la Fuente			Proceso de Transformación	Información del Destino		
Base Datos	Entidad	Atributo		Base Datos	Entidad	Atributo
...	....	...	....	...	...	...
...						
...						

Tabla 3.3 Metadato

El siguiente ejemplo muestra la documentación del metadato para obtener el tipo de garantía y el nivel de aprobación de la base de datos de crédito:

- Información de la fuente de datos:

  - Nombre de la base de datos : Crédito
  - Nombre de la entidad : Cliente, Garantías, Detalle_Garantía
  - Atributo : Tipo de garantía, número de operación,

    detalle de la garantía

- Proceso de transformación:

  Para obtener la garantía de un crédito, se debe realizar un join entre la tabla de prestamos y la tabla de garantías, utilizando como atributo de unión el número de operación. Como resultado de esta operación se obtiene el código de la garantía asociado a la operación de crédito. Con este código se accesa la tabla que contiene el detalle de las características de la garantía.

- Información del destino

  - Nombre de la base de datos : DW Crédito
  - Nombre de la entidad : Dwhp_td_cliente
  - Atributo : Garantía

Como se puede observar en nuestro ejemplo, mostramos parte de la información que puede ser administrada mediante los metadatos. Sin embargo, el metadato debe contener información que describa el acceso, almacenamiento y transformación de los datos, además de:

- Una descripción del modelo de datos.
- Una definición específica del diseño de la base de datos.
- Una definición de los elementos de datos, incluyendo reglas para derivaciones, cálculos y sumarizaciones.

Es a través del metadato que el depósito de datos empieza a ser una herramienta efectiva para toda la empresa. Técnicamente, el repositorio de los metadatos debe ser mantenible y administrable dado que es fundamental en el desarrollo del depósito de datos [GIRA1998], [INGL1997].

Además de la creación del metadato, los requerimientos deben ser analizados desde el punto de vista del volumen de la información. El volumen de la información debe ser estimado, dado que es uno de los principales factores para determinar la clase de tecnología a utilizar. El volumen de los datos impacta al depósito de datos en el tamaño total del depósito y en los tiempos de carga.

Es por ello que se debe realizar un análisis del factor de crecimiento y volatilidad de la información [GIRA1998], [HAMM1996]. Este factor permite definir que tan rápido los datos crecen a través del tiempo y la volatilidad determina el tipo de transacciones que ocurren sobre los datos, mostrando cuales tablas son estáticas y cuales dinámicas.

Aplicación

Una vez realizado el análisis de los requerimientos desde el punto de vista de los datos, se debe proceder a analizar los requerimientos desde la perspectiva de la aplicación. La aplicación es el conjunto mínimo de herramientas disponibles para acceder el contenido del depósito de datos y como estas herramientas van a ser administradas [HAMM1996].

La selección de estas herramientas debe realizarse con base en una clasificación de las necesidades especificas de los grupos de usuarios que utilizarán el depósito de datos [GIRA1998], [HAMM1996]. Desde el punto de vista de necesidades de tecnología, clasificamos los usuarios en tres grandes categorías:

- *Gerentes y ejecutivos de primer nivel:* este tipo de usuarios requieren del contenido analizado del depósito de datos, así como recomendaciones e indicadores empresariales presentados en formas de gráficas y reportes. La información se requiere en términos generales, en forma muy resumida. Necesitan recorrer visualmente la información, por lo que requieren de software muy específico.

- *Gerentes y ejecutivos empresariales:* este grupo de usuarios requieren tener acceso a la información en dos formas. Necesitan los datos en forma de reportes y de gráficas, así como un acceso directo a la información que puede ser incorporando herramientas como hojas de cálculo para un análisis posterior.

- *Analistas empresariales y especialistas en tecnología de la información:* son los principales usuarios del depósito de datos, ellos crean los reportes, las gráficas y las recomendaciones necesarias para la administración. Requieren de herramientas para la minería y el análisis de los datos, requieren acceso tanto a los datos en resumen como en detalle para crear y verificar nuevas hipótesis. Es por ello que el análisis de los requerimientos debe enfocarse en facilitarles

herramientas que les permitan a ellos mismos la construcción de este tipo de consultas.

El establecimiento de estas categorías de usuarios y el análisis de sus necesidades permitirá encontrar el conjunto especifico de herramientas requeridas. Algunos usuarios podrán requerir todas las herramientas mientras que otros usuarios requerirán acceso a un reporte específico.

Tecnología

Otro aspecto que debe ser considerado dentro del proceso de análisis, son los requerimientos de hardware, software y topología de red los cuales deben soportar la implementación del depósito de datos. Una vez analizados los requerimientos desde el punto de vista de los datos, estimados los volúmenes de carga y crecimiento de la información e identificadas las necesidades de los usuarios desde el punto de vista de acceso a la información, se tendrá un panorama más amplio para dimensionar la plataforma tecnológica.

Queda fuera del alcance de esta investigación el detallar los elementos tecnológicos. Sin embargo, se sugiere que esta plataforma sea escalable, portable y mantenible.

Como producto de la etapa de análisis se tienen los requerimientos específicos para cada una de las áreas descritas anteriormente: datos, aplicación y tecnología. El conocimiento derivado de estos aspectos permite un claro entendimiento de los procesos y tareas involucradas en la construcción del depósito de datos.

Dentro de estas tareas se resaltan los procesos de extracción y transformación que deben ser implementados. Esta es una de las áreas en las cuales hay mayor dificultad para encontrar herramientas que permitan administrar y ayudar con estas labores. En nuestro caso la herramienta de extracción y transformación utilizada es el producto “Data

Transformation Services (DTS)", el cual viene con la versión 7.0 del administrador de base de datos SQL Server de Microsoft. La figura 3.6 describe gráficamente el proceso utilizado por nosotros en la construcción del depósito de datos.

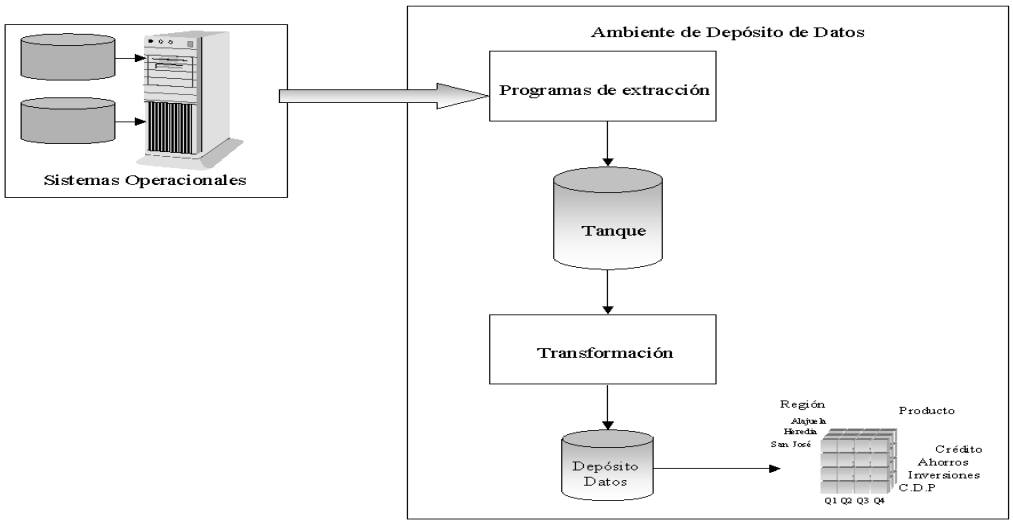


Figura 3.6 Proceso de extracción y transformación

Como se puede observar hemos utilizado programas de extracción para trasladar los datos, sin ninguna alteración, de los sistemas operacionales a una área temporal de trabajo, que llamaremos "tanque". Este esquema nos permite transportar grandes cantidades de información sin sobrecargar los sistemas fuentes con procesos adicionales.

Una vez que la información se encuentre en el "tanque", se procede a aplicar los procesos de transformación, los cuales han sido definidos y descritos de manera general en el metadato. Estos procesos tienen incorporada la lógica que permite hacer la normalización y sumarización de los datos, así como el poblamiento del depósito de datos.

3.3.2 Fase de Diseño

La fase de análisis nos brinda como resultado final el conjunto de especificaciones que permiten derivar en la etapa de diseño el modelo lógico y físico de los datos

[HAMM1996]. La forma para diseñar un depósito de datos es una tarea diferente al proceso utilizado en la construcción de un sistema transaccional. Esta diferencia se acentúa en las tareas requeridas para la obtención de datos.

Es por ello que nuestra fase de diseño se subdivide en los siguientes niveles:

- Diagrama de paquetes de información.
- Esquema estrella.
- Modelo físico de la base de datos.

Cada nivel es esencialmente una versión más refinada o detallada del modelo de datos desarrollado previamente, como se muestra en la figura 3.7. Un modelo de datos es típicamente una representación de la estructura de datos utilizada por algún segmento de la organización y es particularmente útil en la documentación de los datos.

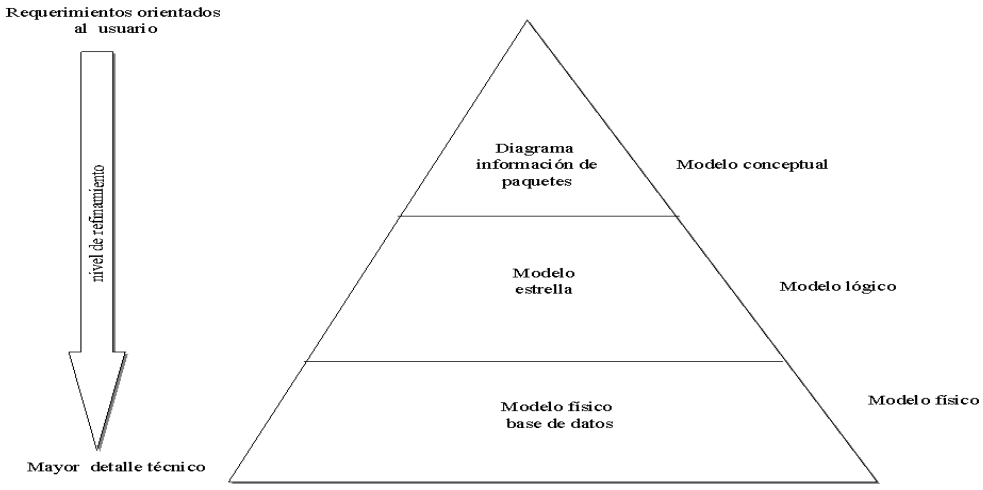


Figura 3.7 Niveles de la fase de diseño.

Estos niveles le permitirá al equipo del proyecto obtener un refinamiento más preciso de los requerimientos de los usuarios y planear mejor los entregables del proyecto. El primer nivel, o modelo conceptual, permite por medio de los diagramas de

paquetes de información modelar y refinar los requerimientos de los usuarios. En el segundo nivel, por medio del esquema estrella, se representa el diseño lógico de la base de datos. Por último, el tercer nivel, se enfoca en el modelo físico de la base de datos.

En los siguientes apartados se explica con mayor detalle estos conceptos.

Diagrama de paquetes de información.

El primer nivel y el más generalizado de nuestro modelo es el diagrama de paquetes de información. Este modelo se enfoca en analizar y estructurar los requerimientos de información de los usuarios, desde el punto de vista de las áreas de negocio y los indicadores que permiten medir la actividad del negocio [HAMM1996]. El diagrama de paquetes de información cumple un propósito muy general, y es proveer una generalización de cómo los datos serán estructurados en el depósito de datos.

Debido a que por naturaleza los datos de la organización son dimensionales [BAUM1996], requerimos de un mecanismo que permita modelar este concepto. El diagrama de paquetes de información provee una técnica para representar la información del usuario en un espacio multidimensional [HAMM1996]. El concepto multidimensional es el termino que se refiere a la información que es definida o accesada a través de varias dimensiones [BAUM1996]. En el mundo de la geometría, una descripción fácil de una entidad multidimensional es un cubo. El cubo tiene tres dimensiones específicas: ancho, alto y profundidad.

Esta técnica se enfoca en la naturaleza dimensional de los datos, es una herramienta que provee una comunicación efectiva entre el equipo técnico y los usuarios del depósito de datos.

Como se menciono previamente, una gran mayoría de los datos de la empresa son dimensionales. Sin embargo, obtener y presentar más de tres dimensiones

tradicionalmente ha sido una tarea dificultosa. El diagrama de paquete de información simplifica esta tarea y permite diseñar y comunicar multidimensionalmente paquetes de información tanto a desarrolladores como usuarios [HAMM1996].

Sin embargo es necesario definir algunos términos que son requeridos para entender mejor esta técnica [BAUM1996], [HAMM1996]:

- *Dimensión:* una dimensión es una propiedad física, tal como tiempo, ubicación o producto. Es un medio fundamental para acceder y presentar la información de la empresa. Típicamente actúa como un índice para identificar datos. Nosotros comúnmente pensamos en un reporte estándar que presenta filas y columnas como dos dimensiones. Un administrador quien esta evaluando el presupuesto puede observar dos dimensiones en una hoja electrónica que contiene las cuentas en la filas y los centros de costos en las columnas. El punto de intersección entre filas y columnas, o una celda, contiene información numérica relevante acerca de un centro de costo especifico y su cuenta.

- *Multidimensional:* este termino se refiere a la información que es definida o accesada por varias dimensiones. Sorprendentemente muchos modelos de la empresa son representados en una vista multidimensional. Un diagrama de paquete de información provee una técnica para modelar la información del usuario en un espacio multidimensional.

- *Medida:* la medida es un mecanismo que mide la información del negocio a través de las dimensiones. Las medidas son típicamente cantidades o capacidades utilizados como comparadores del desempeño de la organización.

- *Categoría:* una categoría es una división específicamente definida en una jerarquía de la dimensión que provee una clasificación detallada del sistema. Este miembro discreto de una dimensión es usado para identificar y aislar

datos específicos. Por ejemplo, la región de Heredia y Guanacaste son categoría dentro de la dimensión de localización. Similarmente, enero y el primer trimestre son categorías dentro de la dimensión de tiempo.

- *Detalle de la Categoría:* es el nivel más bajo de detalle dentro de una dimensión. Por ejemplo, si la dimensión tiempo contiene información acerca de los períodos de tiempo que incluye año, mes y día, día es un detalle de la categoría, el valor de 04/09/2000 es una instancia del detalle de la categoría.

- *Agregación:* en el contexto de depósitos de datos, la agregación se refiere al concepto de sumarizar datos dentro de una jerarquía de dimensiones. Cada dimensión contiene muchos niveles, la información en estos niveles puede ser presentada al usuario como una versión totalizada de los datos. Por ejemplo, en la dimensión de localizaciones todos los distritos pueden ser agrupados para presentar un total por región.

- *"Drill down" y "Drill up***"**. Son técnicas de navegación para que el usuario analice más ampliamente el detalle de la información (down) o agregue los datos en otro nivel de sumarización dentro de una dimensión.

Un diagrama de paquete de información es mostrado en la figura 3.8. En la línea con la etiqueta "NOMBRE DEL PAQUETE" se debe especificar el área del negocio sujeto de análisis. En las columnas se deben nombrar cada una de las dimensiones que se deriven de los requerimientos, estas dimensiones deben ser definidas en términos conocidos por los usuarios: tiempo, localidad, geografía, producto y cliente. Para cada dimensión se debe detallar sus respectivas categorías, hasta el nivel de detalle que se requiera. Por ejemplo en la dimensión de tiempo, se puede requerir detallar la información en las siguientes categorías: año, trimestre, mes, semana, día. El último elemento a documentar en este diagrama son las variables cuantitativas a través de las cuales se mide la actividad del negocio.

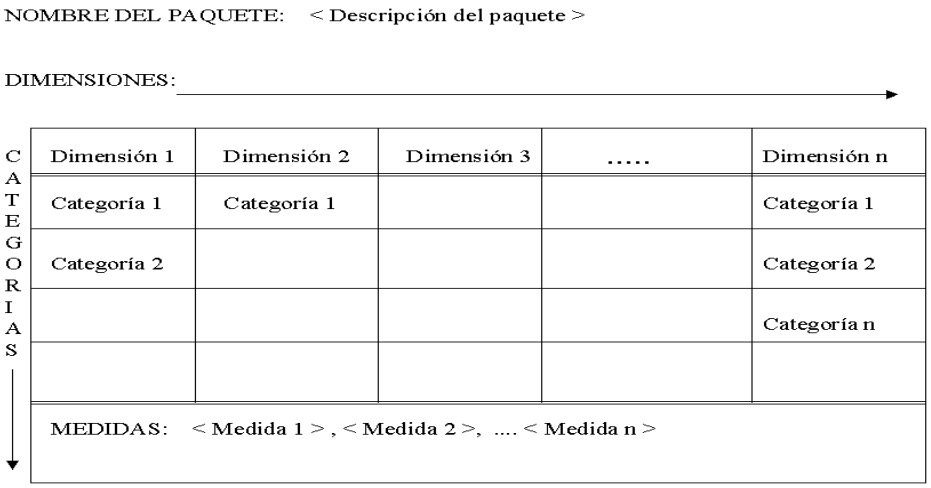


Figura 3.8 Diagrama paquete de información

El diagrama de paquetes de información ayuda en la ejecución de las siguientes tareas [HAMM1996]:

- Define las áreas de temas comunes utilizadas dentro de la organización, tales como tiempo, cliente, geografía y producto.
- Decide como los datos deben ser presentados al usuario del depósito de datos.
- Establece el nivel de detalle (granularidad) de los datos.
- Estima el tamaño del depósito de datos.
- Determina la frecuencia de refrescamiento de los datos dentro del depósito de datos.
- Formula como la información debe ser empaquetada para distribución a los usuarios.
- Define como los datos van a ser accesados, cuales son los puntos de entrada, a donde quiere llegar el usuario y como se navegará en el paquete de información.

Por último, los beneficios derivados de esta técnica se pueden resumir en:

- Permite documentar y estructurar los requerimientos. Es una herramienta que facilita la descripción de cómo el usuario requiere visualizar la información.

- Desarrollo y mantenimiento. Los modelos de datos actúan como un área neutral intermedia entre las aplicación y la base de datos que son desarrolladas. Cuando son construidos apropiadamente, los modelos son independientes de los cambios internos que puedan sufrir las interfaces internas.

El diagrama de paquetes de información cumple un propósito muy general, y es proveer una generalización de cómo los datos son colocados en el depósito de datos, centrándose en el alcance de los requerimientos del usuario y enfocándose en lo que el usuario quiere [GIRA1998], [HAMM1996]. Es un mecanismo efectivo de comunicación entre el equipo técnico y los usuarios, permitiendo identificar cualquier inconsistencia entre los requerimientos y los entregables.

Modelo estrella.

El segundo nivel del proceso diseño es la generación de modelo lógico de datos, para efectos nuestros nos apoyamos en el esquema de estrella, el cual permite visualizar a otro nivel de detalle los requerimientos de los usuarios. La metodología de empaquetamiento provee las bases conceptuales para el esquema estrella y permite su fácil generación [HAMM1996].

El esquema estrella es optimizado para actividades de consulta versus las técnicas tradicionales de modelamiento de bases de datos tales como los esquemas de modelos normalizados [GIRA1998], [HAMM1996]. Los esquemas de modelos normalizados contienen entidades naturales y sus relaciones asociadas. Sin embargo ellos proveen una estructura irregular para los procesos de consulta. En contraste el modelo estrella define las entidades en una forma que soporte la toma de decisiones del negocio, como también que las entidades reflejen los aspectos operacionales importantes de la empresa. Esto es

porque el modelo estrella contiene tres entidades lógicas: dimensión, medida y entidad de detalle de categoría [GIRA1998], [HAMM1996].

Asimismo, este modelo provee un diseño de base de datos que se enfoca en una rápida respuesta a los usuarios del sistema. El diseño que se genera a partir de este esquema no es tan complicado como los diseños de bases de datos tradicionales, lo cual hace que sea más entendible por los usuarios.

Como su nombre lo sugiere el esquema estrella es un paradigma de modelado que tiene un solo objeto en medio conectado con varios objetos de manera radial [HAMM1996], ver la figura 3.9 Esquema estrella.

Un esquema estrella sencillo, consta de una tabla de hechos y varias tablas de dimensión. Una tabla de hechos contiene las mediciones básicas de los negocios, la tabla de dimensión contiene atributos de negocios que se emplean como criterios de búsqueda. El verdadero poder del diseño del esquema estrella es modelar una estructura de datos que permite filtrar las entidades de medidas durante las búsquedas y consultas de los usuarios [HAMM1996].

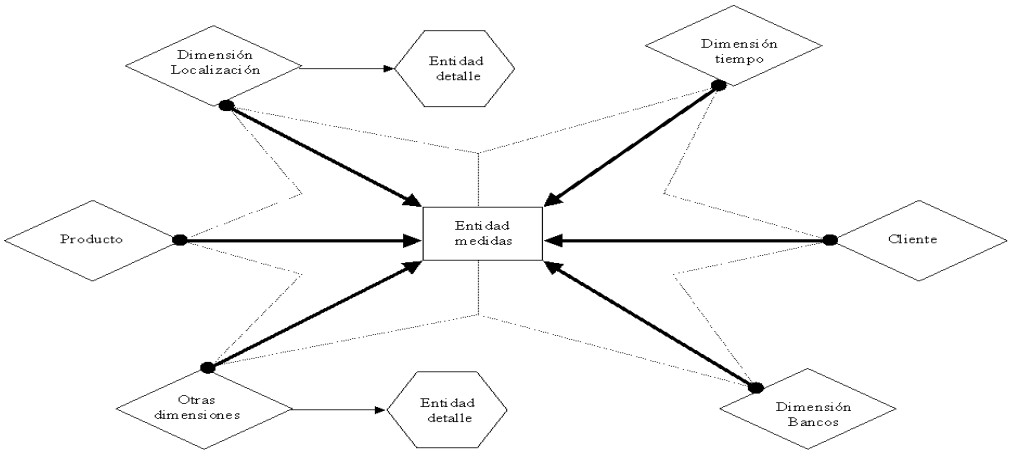


Figura 3.9 Esquema estrella

Un esquema estrella contiene tres tipos de entidades lógicas: *medidas, dimensiones y detalles de categorías* [GIRA1998], [HAMM1996]:

- *Entidad de medidas:* dentro de un esquema estrella, el centro de la estrella y a menudo el foco de la actividad de consulta de los usuarios, es la entidad de medida. Los datos contenidos en esta entidad son llamados "medidas". Las medidas proveen al usuario con datos de cantidad (métricas, indicadores) del negocio. Estas entidades son representadas por un rectángulo.

- *Entidades de dimensión:* son entidades más pequeñas que las entidades de medidas. Las dimensiones y sus datos asociados permiten al usuario visualizar las medidas de una forma más familiar. Estas entidades contienen información de carácter cualitativo. Típicamente son representadas por un rombo.

- *Entidades detalle de categorías:* dentro de un diagrama de paquete de información, cada celda en una dimensión es una categoría y representa un nivel aislado dentro de una dimensión que puede requerir información más detallada. Estas categorías que requieren más detalle son administradas dentro de una entidad de detalle de categoría. Estas entidades tienen elementos que soportan los datos de medidas y proveen más detalle o información cualitativa para asistir el proceso de toma de decisión. Típicamente son representadas por un trapecio.

Consideramos que la técnica del esquema estrella es la más adecuada para modelar lógicamente un depósito de datos. Esta técnica se centra en hacer más eficientes los procesos de consulta en contraposición con los modelos tradicionales, los cuales se preocupan más por las relaciones entre las entidades.

Además, la representación lógica de la base de datos que se deriva del esquema estrella, viene a ser el recurso necesario para el diseño físico del depósito de datos. En el siguiente apartado trataremos este tema.

Modelo físico

El modelo del esquema estrella debe ser utilizado como base del diseño físico de la base de datos y su implementación, porque en él se especifica que datos deben ser incluidos y las relaciones entre las entidades.

En este apartado se describe los aspectos que deben ser considerados para trasladar el esquema estrella al modelo físico de datos, queda fuera del alcance de esta investigación detallar el marco teórico relacionado con en el diseño físico de una base de datos. Sin embargo, resaltamos los conceptos que consideramos de mayor relevancia en esta actividad:

- Definir un estándar de nombramiento de datos.
- Definir las características de las entidades.

Previo a la definición de las características de las entidades físicas, sus relaciones y atributos, se debe adoptar un estándar de nombramiento de los datos que surta información descriptiva acerca de los componentes que serán representados. En términos generales, utilizar palabras completas para el nombramiento de las entidades y atributos, empleando letras en mayúscula, el símbolo de subrayado como separador de palabras y en la medida de los posibles crear listas de dominios para agrupar atributos que representan datos similares.

Una vez establecidos los estándares de nombramiento, se debe transferir las entidades del esquema estrella al modelo físico, definiendo para ello cada una de las características de las entidades. Entre los principales elementos a definir se encuentran: cuales atributos son llaves, rangos validos de datos, tipo y tamaño del dato y las restricciones de integridad.

Estas características pueden ser representadas por medio de la tabla 3.4, donde se muestra un ejemplo de la entidad que contiene las medidas de nuestro modelo.

NOMBRE ENTIDAD:	DMCR_TH_HECHOS			
Nombre del atributo	**Atributo llave**	**Rangos Validos**	**Restricciones de integridad**	**Tipo y tamaño**
CEDAD		De 0 a 99	No puede ser nulo	N(2)
FECHA_CORTE		Fecha de los últimos 5 años	No puede ser nulo	YYYYMMDD
CTIPOLOGIA	Foránea	De 01 a 05	No puede ser nulo	N(2)

Tabla 3.4 Características de las entidades

Además de estas características es importante documentar los volúmenes de información y la frecuencia de actualización de los datos. Lo que permitirá dimensionar los requerimientos de espacio en disco y los tiempos de ejecución de los procesos de extracción y transformación de datos.

Finalmente, como resumen de la propuesta para desarrollar una arquitectura que sirva de procedimiento para el proceso de construcción de un depósito de datos, en la figura 3.10 se muestran gráficamente las fases más importantes de esta metodología.

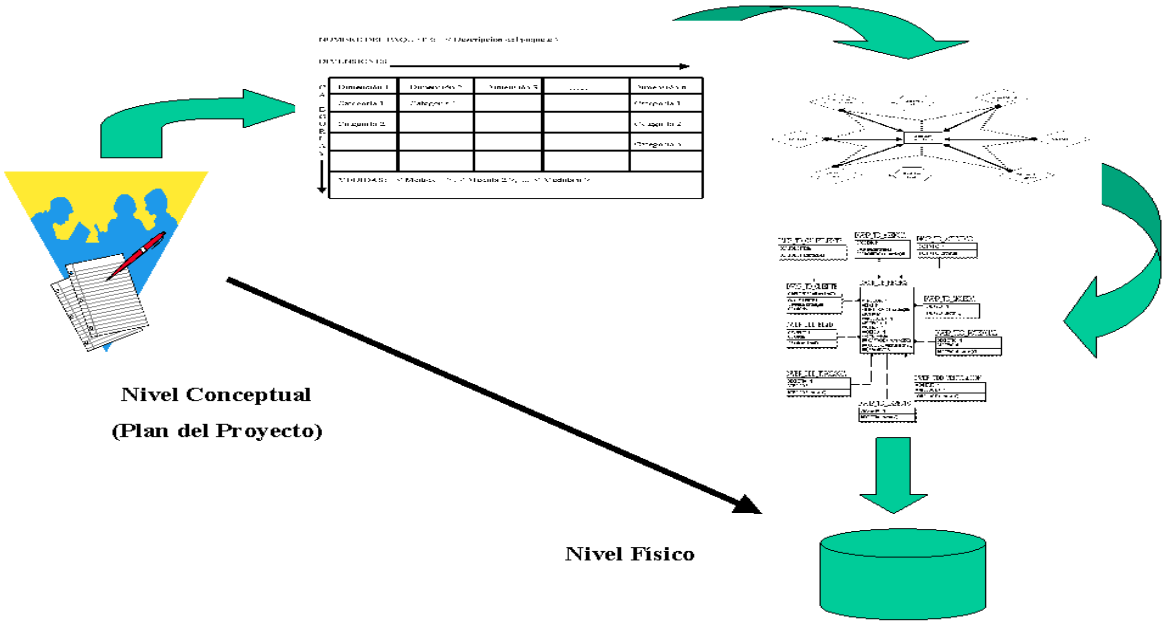


Figura 3.10 Fases principales de la metodología

Lo analizado a través de este capítulo conforma el enfoque a ser utilizado en el proceso de construcción de nuestro depósito de datos. Nótese que esta propuesta permite transformar los requerimientos del usuario obtenidos en la etapa de análisis, en paquetes de información que a su vez son esquematizados por medio del modelo estrella para generar finalmente la especificación física del modelo de datos, secuencia de tareas que finaliza con la construcción del depósito de datos.

Debido a que el alcance principal de nuestro trabajo es el de proponer una metodología que facilite la labor de diseño de un depósito de datos, no se profundiza en las tareas concernientes al proceso de implementación del diseño físico. No obstante se debe definir formalmente una estrategia de implementación que contemple aspectos como:

- Un esquema de prueba que involucre las tareas desde la extracción hasta la consistencia de los datos en el depósito de datos.
- La capacitación a los usuarios.
- Mecanismos para promover la utilización del depósito de datos dentro de la organización.
- Calendarización de las tareas de extracción y transformación.

Finalmente, es importante recordar que un depósito de datos esta basado en un ambiente dinámico en donde los procesos de la empresa y los datos cambian constantemente, es por ello que el proceso de desarrollo nunca estará completo. Razón por la cual se debe continuar desarrollando, componentes adicionales para el depósito de datos, aspecto que con nuestra metodología se facilita.

En el siguiente capítulo se aplica, mediante un caso práctico, cada una de las tareas propuestas en nuestra metodología. Además, se desarrolla una base de datos multidimensional en el que se modelan variables lingüísticas, aplicando para ello el modelo MVL.

CAPÍTULO 4

CASO PRÁCTICO DEL DESARROLLO DE UN DEPÓSITO DE DATOS

"Los ideales son como las estrellas; no puedes tocarlas con tus manos; pero, como los navegantes en el desierto de las aguas, las escoges como tus guías y la sigues para llegar a tu destino."

Carl Schurz

En los capítulos anteriores se presentaron los fundamentos teóricos de la tecnología de Depósitos de Datos y se expusieron los conceptos para modelar variables cuantitativas de forma difusa, de modo que el tomador de decisiones pueda visualizar los datos de una forma más natural. Asimismo, se describieron las actividades que deben realizarse en la construcción de un depósito datos.

Con el objetivo de probar la viabilidad de estos conceptos, este capítulo lo hemos organizado de la siguiente forma: en la primera sección se hace una introducción al planteamiento del problema a partir del cual se deriva una segunda parte en la que se desarrolla un caso práctico que nace como inquietud nuestra y que se fundamenta en dos razones principales: la primera, validar las ideas propuestas en los capítulos anteriores y la segunda, para construir una herramienta que apoye la toma de decisiones en la Dirección Comercial del Banco Nacional. Por razones técnicas, de tiempo y recursos la validación de los conceptos propuestos se realizará mediante la construcción de un *Mercado de Datos*.

El desarrollo de este producto nos permitirá sentar las bases para la creación de un depósito de datos corporativo. Por último se finaliza el capítulo con la generación de una base de datos multidimensional que incorpora variables difusas, poniendo a disposición del tomador de decisiones una herramienta innovadora.

4.1 Planteamiento del problema

Las organizaciones han orientado sus esfuerzos a consolidar la información dispersa en un único repositorio que sirva de base para explotar el proceso de análisis de los datos con la finalidad de ayudar a la toma de decisiones. Este elemento ha permitido el inicio de una nueva actividad, cuyo objetivo es hacer más eficiente los procesos de inferencia en masivos conjuntos de datos.

Se pretende por lo tanto, construir un depósito de datos que facilite al tomador de decisiones analizar y evaluar nuevas tendencias y relaciones entre los clientes y los datos. Con el uso de variables lingüísticas se pone a su disposición una forma más natural para representar y manipular datos cuyos valores están sujetos a un tratamiento impreciso en este tipo de análisis. Por ejemplo, se podría responder a preguntas como: ¿Cuál es la distribución de la colación en los clientes adultos cuyo potencial económico sea alto?.

Con esto la organización dispondrá de una herramienta que le permita orientar sus recursos a los segmentos de clientes que tienen mayor relevancia en el negocio. Se aclara que no se trata de desarrollar una herramienta que permita la segmentación de los clientes, si no de crear una herramienta que permita representar los datos expresados en reglas y valores cuya semántica es más cercana al lenguaje natural del tomador de decisiones.

Para tal efecto se creará un mercado de datos empleando la metodología expuesta en el capítulo 3, utilizando como base cognoscitiva para la aplicación de estos conceptos la Dirección Comercial, la cual es la encargada de establecer los lineamientos a nivel general de cada una de las oficinas que ofrecen productos y servicios a los clientes.

Esta Dirección ha identificado los siguientes objetivos:

- Establecer un sistema de información que permita dar seguimiento y control a la actividad del crédito.
- Brindar apoyo a la comercialización de nuevos productos y servicios.

- Definir lineamientos estratégicos para realizar gestiones personalizadas a los clientes en diversos segmentos por zona geográfica.

4.2. Desarrollo de un caso práctico

De acuerdo a nuestro procedimiento la primera actividad en el proceso de construcción de un depósito de datos es concretar la arquitectura de la planeación, la cual tiene como objetivo principal definir claramente la visión y alcances del proyecto.

Basándonos en el planteamiento del problema, nuestra visión se define de la siguiente forma:

> "*Unificar en un repositorio central la información crediticia a nivel nacional, siendo líderes en la aplicación e implementación de la tecnología Depósitos de Datos, con el objetivo de proveer información de crédito a las respectivas áreas en forma ágil, rápida, oportuna y expedita para una eficiente y productiva toma de decisiones a través de los diferentes niveles de decisión".*

Este proyecto permitirá mostrar y analizar en forma resumida la información crediticia que se encuentra localizada en las diferentes fuentes o base de datos de la Institución. Para ello se requiere contar con una base de datos actualizada y cargada mensualmente, cuyas fuentes principales de información son los sistemas de Crédito, Tarjeta de Crédito y Sobregiros.

El manejo de proyecciones, minería de datos, análisis de sensibilidad y otros de este tipo no están incluidos como alcance del proyecto. La información de crédito requerida es la siguiente:

- Información de crédito a través de su fecha de saldo, formalización, amortización y vencimiento; jerarquizada hasta el nivel mensual.
- Información de crédito por Región u Oficina a nivel nacional, hasta el nivel de cliente.

- Información de crédito por tipo de cartera: comercial, hipotecario, junta rural, mandato legal.
- Información de crédito por moneda: colones, dólares, euros.
- Información por estado del crédito, hasta el rango de atraso del principal y/o interés.
- Información de crédito por actividad económica.
- Información de crédito por composición de la cartera: activa e inactiva.
- Información de crédito por grupos de interés económico.
- Información de crédito por producto.
- Información de crédito medido a través de:
    - Saldo en unidades monetarias
    - Tasas de colocación
    - Plazos de Colocación

Debido a que este es el primer proyecto de esta clase en la Organización, el proceso de construcción y desarrollo es inmaduro, razón por la cual se debe iniciar las labores con un equipo base que adquirirá experiencia y madurez conforme se avance en el proyecto. Sin embargo, como parte de las tareas de planeación se definieron los siguientes perfiles para el equipo de trabajo:

- *Administrador del proyecto:* debe ser una persona con experiencia en el negocio, con capacidad de liderar proyectos a nivel macro y con amplios conocimientos de las necesidades del negocio. Debe tener autoridad para tomar decisiones y con conocimientos del manejo de sistemas computacionales.

- *Analista del negocio*: persona con conocimiento de las necesidades del negocio y con la capacidad de liderar el proyecto en el día a día, además de la autoridad necesaria para tomar decisiones.

- *Arquitectos*: personas con el grado de Ingeniero de Sistemas, capacitados en la Administración de Base de Datos y aspectos básicos del negocio.

Como parte de las labores de planeación se considero oportuno analizar y documentar los principales riesgos a los cuales el proyecto podría estar expuesto. Para ello se utilizo la siguiente tabla, donde se describe el tipo de riesgo, su probabilidad de ocurrencia y el impacto que este podría tener en el proyecto:

Tipo de Riesgo	Probabilidad de Ocurrencia	Impacto sobre el Proyecto
Disponibilidad Equipo de Desarrollo	5%	10%
Disponibilidad de Infraestructura	0%	0%
Confiabilidad de la Información	66%	100%
* Saldos Cartera vs. Contabilidad	1%	100%
* Clasificación de Cartera	40%	100%
* Clasificación de Riesgo	25%	100%
* Saldo Interés de Mora	0%	0%
No identificar un cliente único	75%	100%
Conectividad de sistemas	0%	0%
Involucramiento de otras áreas	0%	0%

Tabla 4.1 Principales riesgos del proyecto

El desarrollo de las actividades anteriores, como lo son la visión, alcances y objetivos del proyecto, además de la estructuración del equipo de trabajo y la identificación de los principales riesgos a los cuales podría verse expuesto el proyecto, nos permitieron tener un panorama más claro de las necesidades de los usuarios y sus expectativas. Asimismo, nos permitió establecer las reglas que regirán todo el proceso de la construcción del depósito de datos así como dimensionar el tiempo y los recursos requeridos en cada una de las tareas.

Siguiendo los pasos de nuestro procedimiento en la construcción del depósito de datos, la siguiente actividad a desarrollar es la arquitectura actual, aspecto que ya fue detallado en el apartado 3.2 de nuestra tesis. En ese apartado describimos por medio de la arquitectura de aplicación las funciones de negocios requeridas dentro del alcance del proyecto, además, por medio de la arquitectura de datos se mostró a nivel macro el modelo de datos con las relaciones entre las diferentes entidades. Mediante la arquitectura de tecnología se muestra a nivel general los diferentes componentes tecnológicos empleados por las diferentes aplicaciones.

Como fase final de nuestro procedimiento, en los siguientes párrafos se procede a detallar los aspectos que consideramos más relevantes dentro de la arquitectura del ciclo de vida, la cual debe ser iniciada con un claro entendimiento de las necesidades de las áreas funcionales de la organización contempladas dentro del alcance del proyecto. Este análisis generó como producto el refinamiento de los requerimientos del usuario, desde la perspectiva de la granularidad y sus dimensiones:

- *Granularidad**:***

  Este aspecto brinda el nivel de detalle de la información requerida por el usuario.

  - Fecha de Saldo: fecha de corte de la información de la operación de crédito hasta nivel de mes.

  - Fecha de Formalización: fecha cuando se activo contablemente la operación de crédito. Jerarquizada hasta nivel de mes.

  - Fecha de Vencimiento: fecha en la que está pactada la cancelación de la operación de crédito. Jerarquizada hasta nivel de mes.

  - Fecha de Servicio de Interés: fecha hasta que está cubierto el pago de interés normal de una cuota (esta quiere decir cuando se pago la ultima porción de interés de una cuota). Jerarquizada hasta nivel de mes.

  - Fecha de Amortización: fecha hasta que está cubierto el pago de la amortización del principal (pago de la cuota). Jerarquizada hasta nivel de mes.

  - Fecha de Documento: fecha en la que se firma el contrato legal de una operación de crédito. Jerarquizada hasta nivel de mes.

  - Regiones: información de operaciones de crédito a nivel regional y por oficina.

  - Tipo de Cliente: clasificación del cliente a quien se otorga el crédito.

    Física : persona natural con cédula de identidad o pasaporte del extranjero.
    Jurídica : persona jurídica inscritas en el Registro Público Mercantil.

  - Tipo de Cartera: Clasificación Contable donde se ubica el crédito: Comercial e Hipotecario.

  - Moneda: tipo de unidad monetaria en la que se otorga una operación de crédito.

Colones : Moneda local.
Dólares : Moneda extranjera.
Euros : Moneda extranjera.

- Actividad Económica: hacia donde se asignan los recursos de las operaciones de crédito.

  01 Agricultura y silvicultura
  02 Ganadería o pesca
  03 Industria de manufactura y extracción
  04 Electricidad, gas, agua, servicios sanitarios y otras fuentes de energía
  05 Comercio
  06 Servicios
  07 Transporte y comunicaciones
  08 Depósitos y almacenamientos
  09 Vivienda
  10 Construcción
  11 Consumo
  12 Turismo

- Composición de Cartera: operación de crédito que generan o no ingreso financiero: Activa e Inactiva.
- Sistemas: diferentes fuentes de información de operaciones crediticias de las oficinas.

- *Dimensiones:*

  Las siguientes son las dimensiones definidas con el objetivo de estructurar los datos.

  - Fechas:

    Año
    Semestre
    Trimestre
    Mes

  - Regiones

    Institución
    Subregiones
    Oficinas
    Ejecutivo/Responsable
    Cliente

- Tipo de Cliente

  Físico
  Jurídico

- Tipo de Cartera

  Comercial
  Hipotecaria

- Moneda

  Colones
  Dólares
  Euros

- Estado de Crédito

  Vigente
  Vencido
  Cobro Judicial

  Rango de Atraso

  - 1-15 días
  - 15-30 días
  - 30-45 días
  - 45-60 días
  - 60-90 días
  - 90-120 días
  - 120-180 días
  - 180-360 días
  - Más de 360 días
  - Más de 60 días
  - Más de 90 días

- Actividad Económica

  Tipo de Actividad
  Tipo de Rubro / Sección
  Subtipo de Rubro/Sección

- Composición de Cartera

  Activa
  Inactiva

- *Medidas Cuantitativas*

  Con respecto al Crédito

  - Monto Inicial
  - Saldo de Capital

- Saldo de Interés
- Saldo Total
- Multa por Atraso Más de 10 días
- Gastos Judiciales.
- Honorarios.
- Gastos por Póliza.
- Comisión.
- Gastos Estudios Interno.
- Tasa de Interés
- Plazo
- Número de Operaciones
- Número de Clientes
- Número de Arreglos de Pago
- Número de Incumplimiento de Arreglos de Pago
- Número de Cuotas
- Número de Días de Atraso Capital
- Número de Días de Atraso Interés

Con respecto a la Cuota

- Saldo de Ajuste
- Saldo Cuota
- Saldo Cuota Amortización
- Saldo Cuota Interés
- Interés Moratorio

Siguiendo los pasos de nuestro procedimiento la siguiente actividad consiste en crear el metadato que permitirá documentar el origen de los datos y los algoritmos de transformación para lograr obtener la información requerida por el usuario en la etapa anterior. Remitimos al lector al apéndice C, donde mostramos por medio de un ejemplo un metadato, correspondiente a uno de los sistemas que genera información para el depósito de datos.

La conclusión del metadato es la tarea final del proceso de análisis, convirtiéndose en la fuente principal de información de la próxima fase de nuestro procedimiento. Fase que consiste en la generación de los paquetes de información, para que por medio de ellos se derive el modelo lógico y físico del depósito de datos, conceptos descritos en el apartado 3.3.2.

La siguiente actividad de nuestro proceso es la fase de diseño, en la cual por medio de los diagramas de paquetes de información se refinan y modelan multidimensionalmente los requerimientos del usuario. Para que en una segunda fase se genere el modelo lógico de la base de datos, utilizando como medio de representación el esquema estrella. Por último, se construye el modelo físico de la base de datos.

Continuando con el desarrollo de nuestro caso práctico, mostramos en la siguiente figura un ejemplo de un paquete de información.

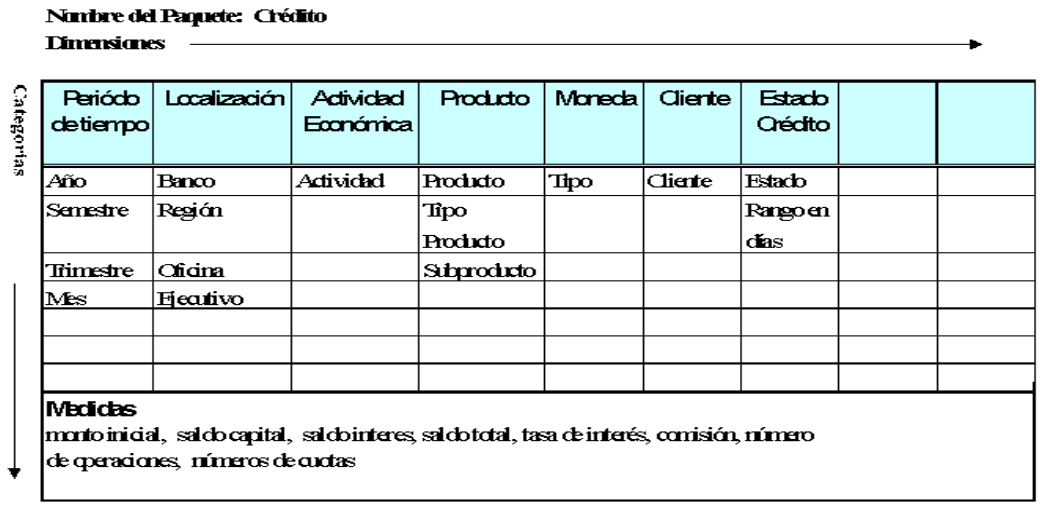


Figura 4.1 Diagrama de paquete de información del crédito

Mediante este paquete se modela como los requerimientos del usuario serán estructurados en el depósito de datos. Por ejemplo, la dimensión *periodo de tiempo* se refiere a todos aquellos elementos relacionados con los requerimientos del usuario asociados con las fechas, las cuales serán jerarquizadas en termino de año, semestre y mes. Lo cual indica que el usuario podrá visualizar la información crediticia detallada hasta un nivel de mes. Asimismo, la columna *Localización* representa una dimensión de localidad que establece el nivel de detalle o granularidad de los datos, desde el nivel Institucional hasta el ejecutivo responsable del crédito del cliente.

Además, por medio del paquete de información se representan las variables cuantitativas o indicadores a través de los cuales se medirá la actividad crediticia. Por

ejemplo, el usuario podrá visualizar el saldo y el monto inicial de las operaciones agrupadas por período o región, facilitando al tomador de decisiones el análisis de la información.

El diagrama de paquete de información presenta la definición conceptual de la información requerida por el usuario, además, de ser una efectiva herramienta de comunicación. Una vez finalizado el modelamiento de los requerimientos del usuario por medio de los paquetes de información, se procede a la generación del modelo lógico de los datos. En la figura 4.2 se muestra el esquema estrella derivada de la información contenida en los paquetes de información.

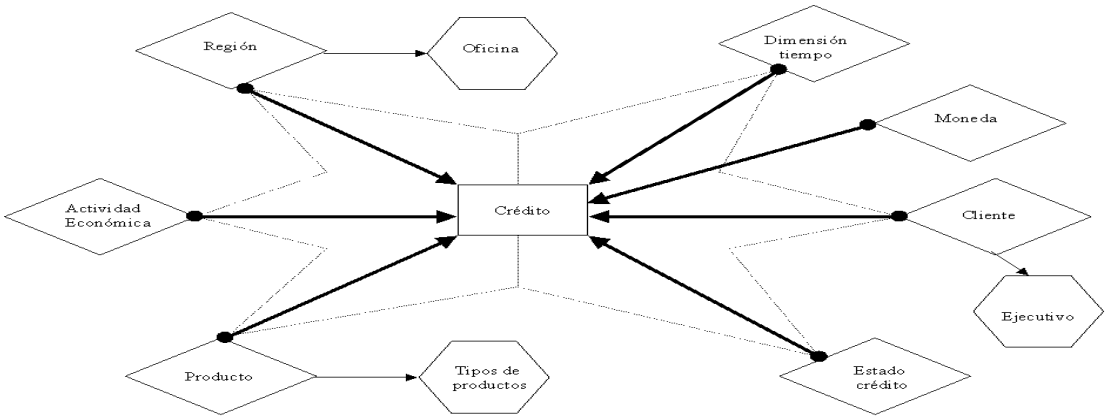


Figura 4.2 Esquema estrella para la actividad crediticia

El esquema estrella esta compuesto por tres tipos de entidades lógicas: medidas, dimensiones y detalle de categorías. La entidad de medidas estará conformada por las diferentes variables cuantitativas descritas en el paquete de información como medidas. Cada una de las dimensiones del paquete será representada en el modelo lógico por medio de una entidad de dimensión. Las entidades de detalle representan aquellas dimensiones que requieren un mayor nivel de detalle.

Cada diagrama de paquete de información representa una estrella completa. En la figura 4.3 se muestra paso a paso como trasladar el paquete de información a la estrella. En el primer paso, punto 1, se ubica en el centro de la estrella la entidad de medida, que

será nombrada como la tabla de hechos en el modelo físico, los atributos de esta entidad estarán conformados por el nivel más bajo de cada una de las categorías y las medidas descritas en el paquete.

El segundo paso, punto 2, consistirá en trasladar cada entidad de dimensión del diagrama de paquete de información a la periferia de la estrella, simbolizando de esta forma los puntos de la estrella. De requerir mayor detalle sobre algunos de los aspectos de una dimensión, por ejemplo oficina, ésta será especificada en el modelo estrella como una entidad de detalle de categoría, como se muestra en el punto 3.

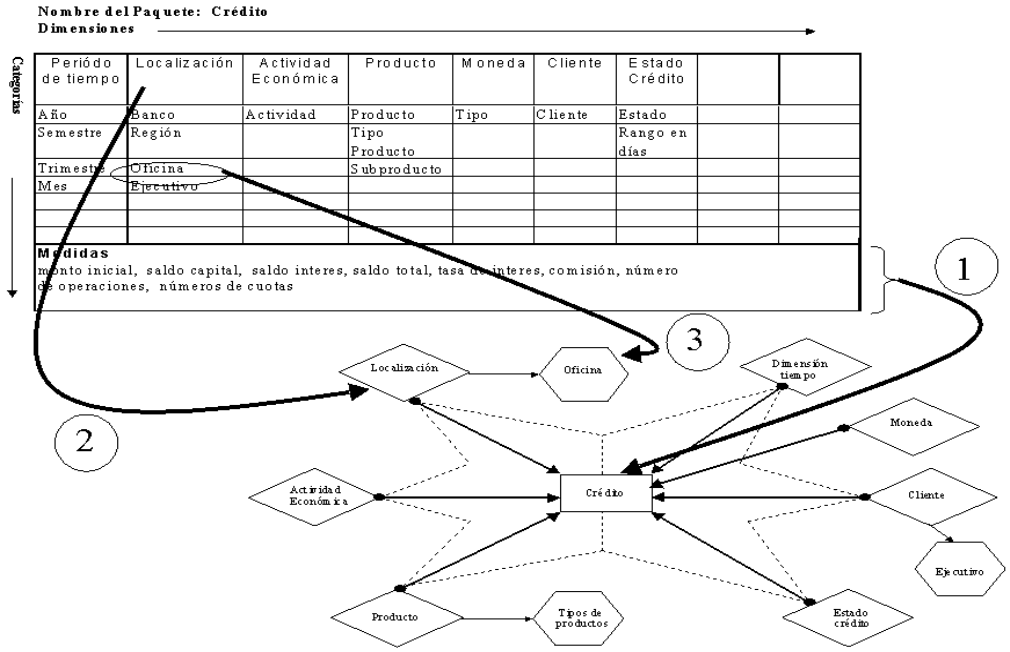


Figura 4.3 Traslado del paquete de información a la estrella

Una vez finalizado el modelo lógico, la siguiente tarea consiste en la generación del modelo físico de la base de datos, para ello se debe generar inicialmente la tabla de hechos, la cual contendrá como atributos las medidas y el último nivel de detalle de cada una de las dimensiones. En la figura 4.4, se muestra gráficamente el proceso para crear la tabla de hechos:

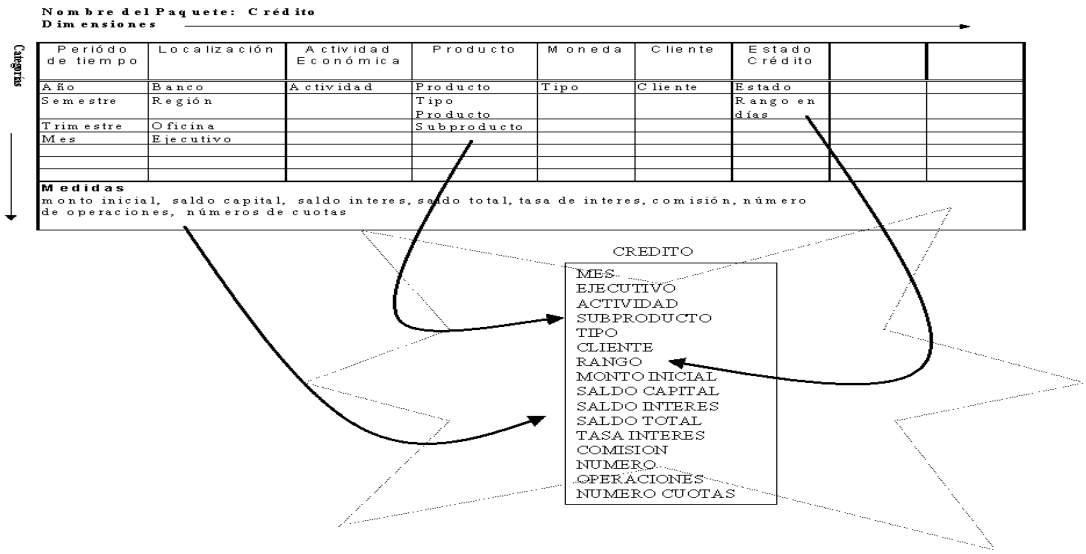


Figura 4.4 Creación de la tabla de hechos

Seguidamente se procede a modelar cada una de las restantes entidades del modelo físico. En la figura 4.5, se ilustra este proceso:

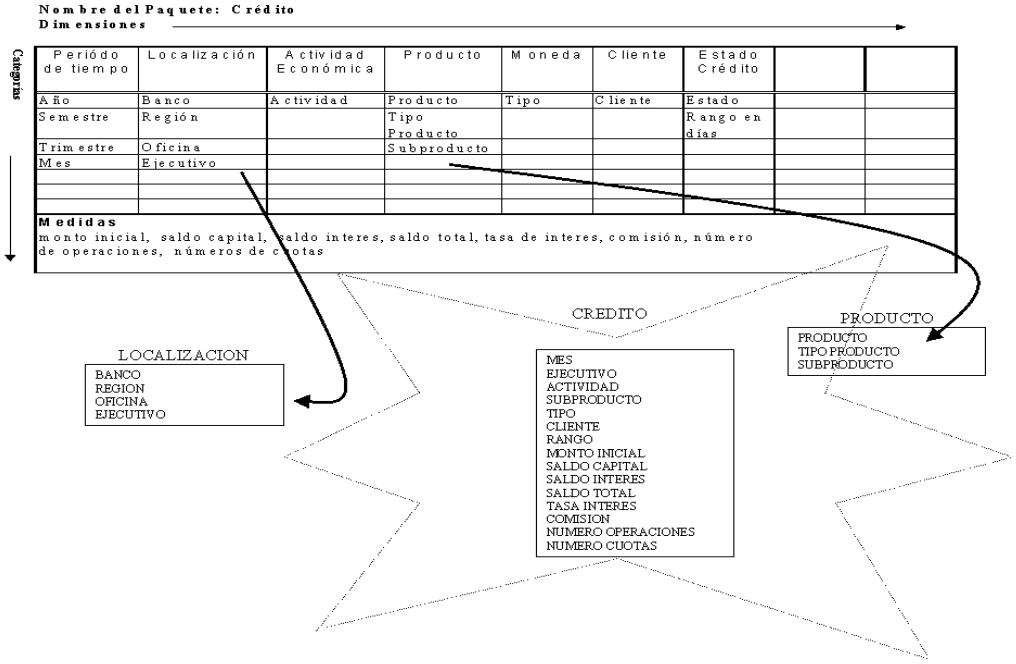


Figura 4.5 Generación de entidades

Para efectos de nuestro ejemplo, hemos mostrado el proceso de construcción de la entidad localización y producto. Los atributos de cada una de estas entidades corresponderán a las categorías de cada una de las dimensiones.

Las dimensiones están colocadas en los puntos de la estrella y tienen una relación hacia el interior de la estrella. La relación entre las dimensiones y la entidad de medidas es de uno a muchos; una instancia de la entidad de dimensión se relaciona con muchas instancias de la entidad de medidas. Estas entidades de dimensión por lo general son de los elementos más desnormalizados del modelo.

El modelo estrella simplifica el modelo lógico de datos organizando las entidades de datos en una manera más optima para el procesamiento analítico. En resumen, el modelo físico estará conformado por una entidad central (medidas) y un conjunto subordinado de entidades de dimensión para efectos de navegación.

Como punto final de nuestro caso práctico, en lo que corresponde a la aplicación del procedimiento descrito en el capítulo 3, mostramos en la figura 4.6 el modelo físico utilizado para la construcción del depósito de datos. Remitimos al lector al apéndice D, donde se muestra con mayor detalle las características físicas de los atributos de las entidades y los estándares de nombramiento de los datos e entidades.

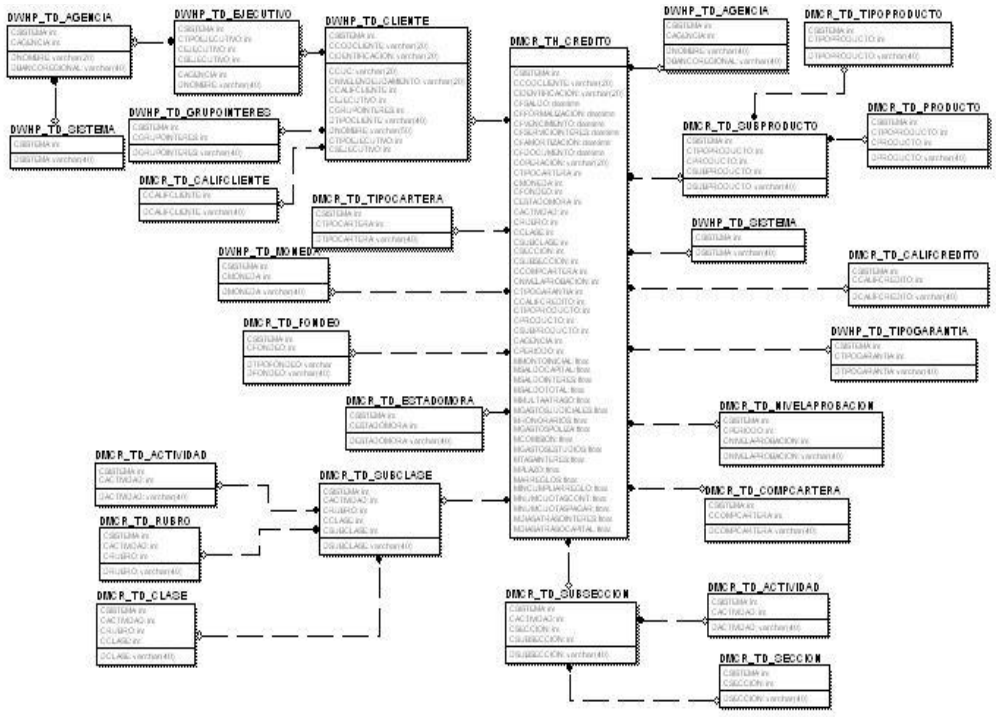


Figura 4.6 Modelo físico

Las siguientes actividades en el proceso de construcción del depósito de datos son: la generación de la base de datos derivada a partir del modelo físico, la extracción, poblamiento y transformación. Como se describe en el capítulo 3 (figura 3.6), para agilizar el proceso de extracción y transformación hemos utilizado una base de datos temporal de trabajo denominada “tanque”.

Para la creación del “tanque”, seleccionamos el Administrador de Bases de Datos Relacional SQL Server 7.0 de Microsoft, los aspectos de programación necesarios para la extracción y transformación se realizaron con Transact-Sql y Data Transformation Services (DTS) de Microsoft. En el apéndice E, se presentan los principales programas utilizados para la extracción y poblamiento de esta base de datos.

Para almacenar la información unificada y transformada del tanque, se crea la base de datos definida en el modelo físico mediante SQL Server 7.0,. El modelo de esta base de datos, responde a las necesidades planteadas por la Organización en la definición de los respectivos requerimientos, los cuales han sido derivados a partir de nuestro procedimiento. En el apéndice F, se muestran los principales programas utilizados en el proceso de transformación y carga del mercado de datos.

Finalmente, se utilizaron los productos Impromptu, Transformer y Power Play de la empresa Cognos para generar y acceder la base de datos multidimensional. Queda fuera del alcance de nuestra tesis brindar una explicación exhaustiva de cómo utilizar estas herramientas, sin embargo, consideramos oportuno mencionar las generalidades más importantes de cada una de ellas.

Con la herramienta Impromptu se genera el catálogo de todos los elementos de datos que serán utilizados en las consultas de los usuarios. Este catálogo permite definir las estrategias de unión a ser utilizadas entre las diferentes tablas. Mediante la herramienta Transformer se define la forma de cómo se presentarán los datos al usuario, además, de generar la base multidimensional. Por medio de Power Play el usuario final visualiza y consulta la información almacenada en el cubo.

Como resultado del procedimiento descrito a lo largo del caso práctico, en donde originalmente nos planteamos crear un mercado de datos para la Dirección Comercial que permitiera al usuario el análisis de la actividad de crédito, procedemos a mostrar el producto final que estará a disposición del tomador de decisiones.

En la figura 4.7 se ilustra como el concepto teórico de la base multidimensional desarrollado a través de este capítulo, se plasman en una herramienta gráfica que permite al usuario el análisis de los diferentes indicadores del negocio, indicadores que fueron desarrollados a partir de sus requerimientos. Como mencionamos la herramienta para manipular y analizar la información contenida en el Cubo es Power Play, la cual tiene dos áreas principales: el área de dimensiones y medidas, y el área de trabajo.

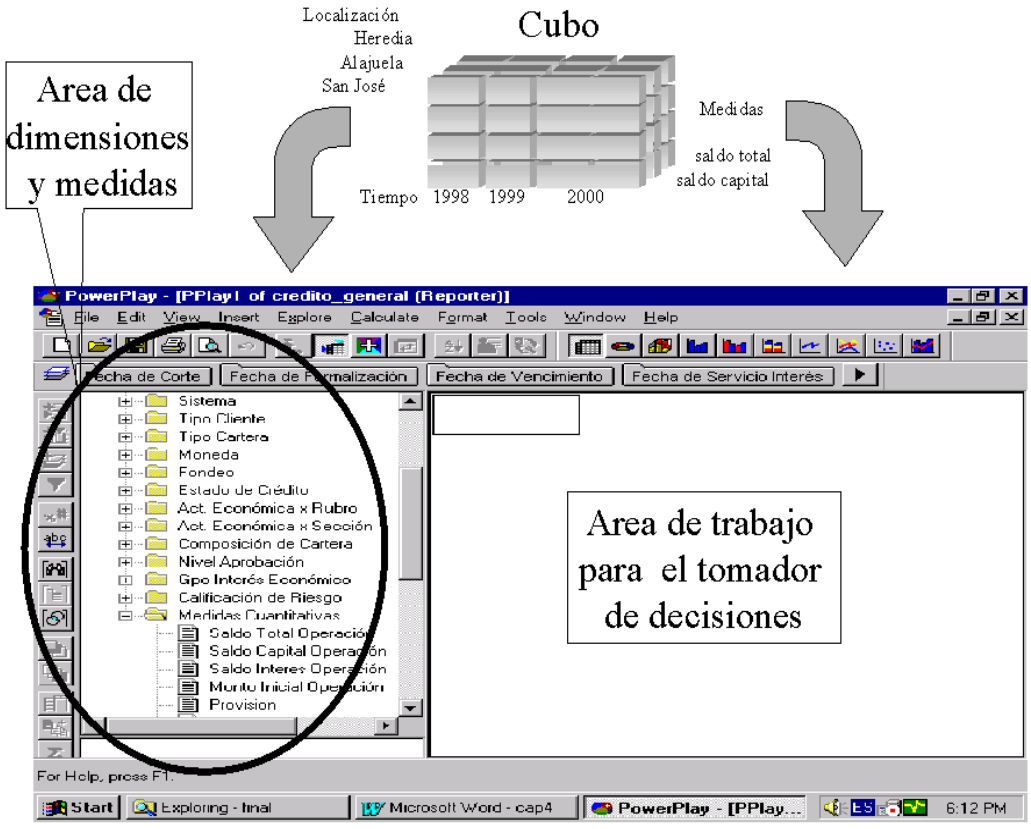


Figura 4.7 Herramienta Power Play

El área de medidas y dimensiones agrupa todos elementos de datos definidos por el usuario desde la perspectiva del tema, granularidad y dimensión. Asimismo, en esta área se presentan las diferentes medidas o indicadores a través de los cuales el tomador de decisiones mide la actividad del negocio. El área de trabajo será la zona en la cual el usuario correlaciona las medidas y dimensiones de acuerdo al análisis específico que este realizando. Por ejemplo, en la figura 4.8 se muestra por medio de un histograma el saldo total de las operaciones por moneda para los diferentes períodos de tiempo.

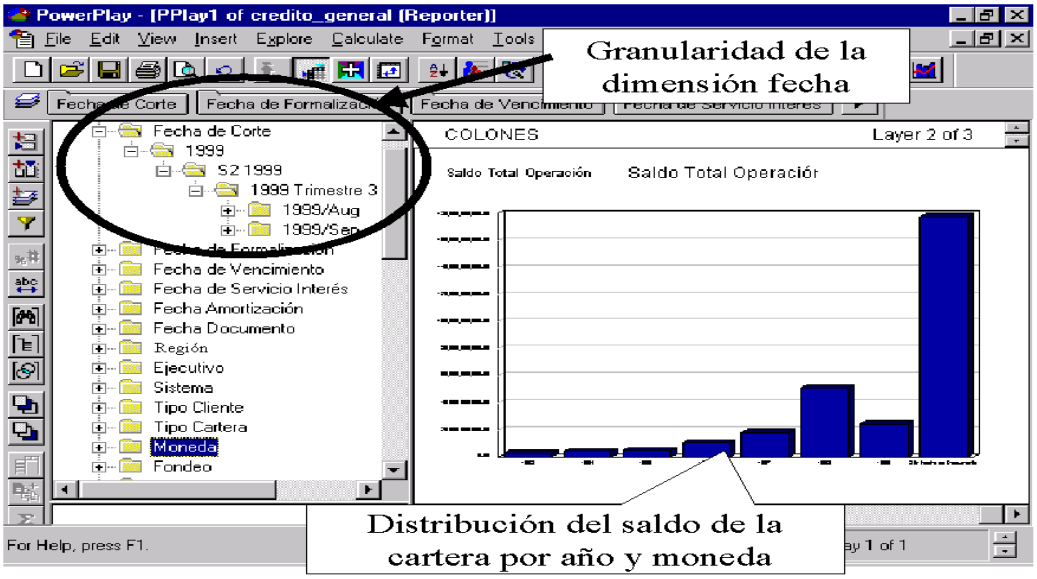


Figura 4.8 Análisis del saldo de la cartera

Como se puede observar, la dimensión de “fecha” ha sido expandida con el objetivo de mostrar su nivel de granularidad. En este caso el usuario puede visualizar la información del “saldo de la cartera” hasta un nivel de detalle mensual.

Por medio de la figura 4.9 se muestra otro ejemplo de cómo el tomador de decisiones puede visualizar el “saldo de la cartera” desde la perspectiva del tipo de cliente, es decir, podrá analizar el comportamiento de esta medida en función de los clientes físicos o jurídicos.

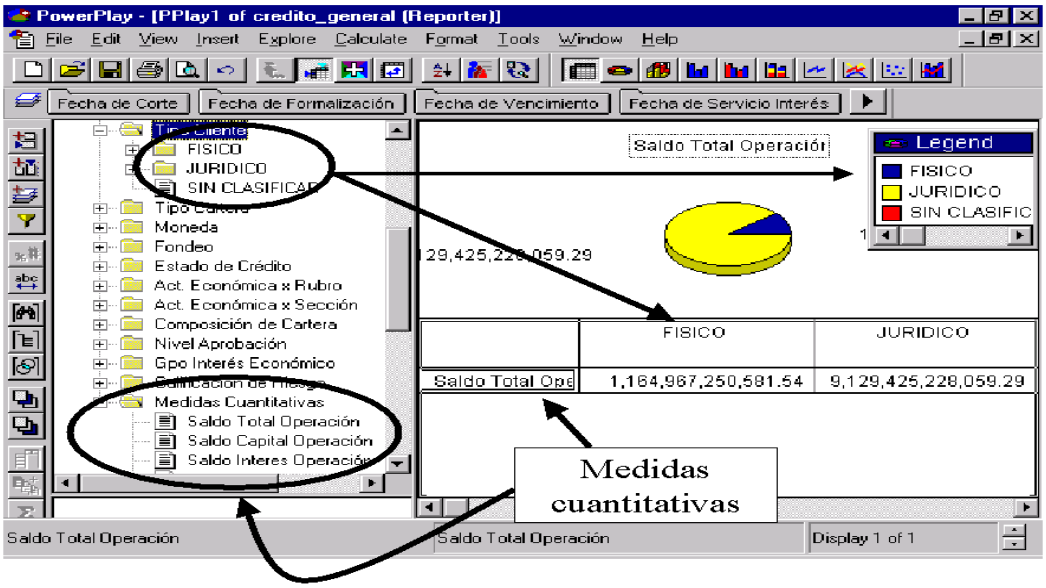


Figura 4.9 Distribución del saldo de la cartera por tipo de cliente

Con estos dos ejemplos pretendemos mostrar como de una manera muy intuitiva el tomador de decisiones podrá hacer uso de todos los recursos contemplados en la base multidimensional, la cual ha sido derivada por la aplicación de cada una de las tareas y actividades contempladas en el procedimiento para la construcción de un depósito de datos.

Sin embargo, debido a que existe una gran cantidad y variedad de datos cuya naturaleza no permite ser formulados de forma precisa, en el siguiente apartado procedemos a desarrollar una base multidimensional que incorpore variables lingüísticas, con el objetivo de permitir al tomador de decisiones visualizar los datos de forma más natural.

4.3 Base Multidimensional Difusa

Puesto que el objetivo principal de este trabajo es el desarrollo de un procedimiento para la construcción de un depósito de datos que incorpore variables lingüísticas, modelándolas con la noción de grado de membresía, para permitir al tomador de decisiones visualizar los datos de forma más natural, procedemos en esta

sección a describir como agregar a la base de datos multidimensional, las variables lingüísticas que permitirán brindar el valor agregado a nuestro trabajo de investigación.

Para tal efecto, en primera instancia, se formularan las variables que serán incluidas en la base de datos multidimensional y que el procedimiento para su declaración fue descrito a lo largo del capítulo 2, bajo el esquema del modelo MVL.

Posteriormente se ilustraran las principales fases de nuestro procedimiento que describen y justifican los lineamientos para la creación de esta nueva base de datos. Por último se detallará su funcionalidad a través de ejemplos gráficos que nos permitirán el análisis de los resultados.

4.3.1 Definición de variables lingüísticas

La idea de definir estas variables lingüísticas se fundamentan en dos razones principales: la primera, para validar las ideas propuestas en los capítulos anteriores y la segunda, para construir una herramienta que apoye a la toma de decisiones para la Dirección Comercial. Además, el desarrollo de este nuevo modelo de datos siente las bases para la creación de un nuevo producto que llegue a ser comerciable.

El objetivo es relacionar variables que son cualitativas con variables que son cuantitativas pero que se puedan difusificar, de modo que se puedan hacer consultas sobre ellas. Estas consultas al plantearse como predicados de primer orden, la lógica difusa las puede manipular por medio de las variables lingüísticas, brindando resultados significativos que con los métodos tradicionales no podrían darse con facilidad.

Debido a que en el ámbito bancario se esta gestando un cambio profundo en la forma de analizar y utilizar la información, hemos considerado conveniente incorporar al modelo de datos las siguientes variables tipología, vinculación, potencial y edad, las cuales permitirán tener un mayor entendimiento del cliente. Variables cuya medición se deriva a partir de valores puntuales y subjetivos.

Estas variables permitirán al tomador de decisiones disponer de un modelo más rico semánticamente, facilitando de esta forma su labor de análisis al responder a preguntas como:

- ¿Dónde estamos actuando?
- ¿Con quien trabajamos?
- ¿Quiénes son los clientes?
- ¿Qué aportan al volumen del negocio?
- ¿Está correctamente definido el mercado geográfico del Banco?
- ¿Qué clientes debemos retener y a qué clientes debemos vincular?
- ¿Qué potencial tiene un cliente?

A continuación se define el significado de las variables consideradas y se justifica la utilización de cada una de ellas:

- *Tipología*

  Este concepto es la categorización de los clientes con el fin de asociarlos a modelos de conducta similares, es modelado con base a su edad, permitiendo analizar su comportamiento a fin de identificar su capacidad de endeudamiento; clientes de alta relación, clientes con riesgos y oportunidades y clientes bien relacionados. Los valores de esta variable lingüística serán:

  - Menores de edad
  - Primer empleo
  - Hipotecarios
  - Inversión
  - Previsión
  - Conservadores

  De este modo el tomador de decisiones podrá visualizar los clientes con riesgo y oportunidades, o sea, valores de: menor de edad, primer empleo, e

hipotecario. Asimismo, analizará los clientes con alta relación: valores hipotecario e inversión. Estos criterios le permitirán orientar la asignación de nuevos créditos o productos a estos segmentos.

- *Vinculación*

  Con el concepto de vinculación se modela que tan ligado a la Institución se encuentra el cliente. Los clientes con poca vinculación se les deberán incrementar esta relación, a los de alta vinculación se le debe fortalecer y mantener su relación. El tipo y número de productos y servicios que el cliente posea con la Institución son empleados para determinar su vinculación. Los valores de esta variable lingüística serán:

  - Alta vinculación
  - Media vinculación
  - Baja vinculación

- *Potencial*

  El potencial es el nivel de capacidad de un cliente para realizar negocios financieros, se determina por medio del saldo promedio mensual de sus captaciones. Para efectos nuestros este valor se obtiene de la suma de los saldos de las cuentas corrientes y de ahorros que el cliente tenga con la Institución. Los valores de esta variable lingüística serán:

  - Potencial alto
  - Potencial medio/alto
  - Potencial bajo/medio
  - Potencial bajo

Esta variable le permitirá al tomador de decisiones conocer la distribución de los montos captados por grupo de cliente, de forma que se orienten las políticas para la venta de productos y servicios.

- *Edad*

Los valores lingüísticos de la variable edad, se determinan con base en la edad de cliente. Se modela esta variable para segmentar los clientes desde una perspectiva difusa, con el objetivo de enriquecer semánticamente el modelo. Los valores de esta variable lingüística serán:

- Viejo
- Adulto
- Joven

El uso de estas variables tiene como objetivo clasificar los clientes y de asociarlos a modelos de comportamiento y consumo similares, permitiendo al tomador de decisiones construir grupos homogéneos de clientes por su tipología, potencial, y la de analizar en profundidad su forma de relación desde el punto de vista de su vinculación. De esta forma se podrá analizar la información de los clientes más importantes de cada grupo o tipología con el fin de disponer de la riqueza adecuada de cada análisis.

Los valores de las variables cuantitativas son traducidos a las etiquetas lingüísticas por medio de los criterios y de los umbrales de aceptación de los expertos, definidos de la misma forma como se ilustro en la tabla 2.7 valores difusos del capitulo 2.

Es muy factible que algunas de los atributos considerados, carezcan de importancia en otros modelos y que otros atributos no considerados, tengan un gran peso en este modelo. Sin embargo, las variables propuestas constituyen un primer intento por demostrar la utilidad de la lógica difusa y específicamente las variable lingüísticas para el

modelamiento de una base multidimensional, empleada como herramienta de apoyo a la toma de decisiones en ambientes tan complejos e inciertos como lo es la banca.

Con esto se concluye la aplicación de la incorporación de las variables difusas en el modelo MVL. En el siguiente punto se procede a ilustrar el resultado final de la generación de la base datos multidimensional difusa.

4.3.2 Creación de la base de datos multidimensional difusa

Como ya se menciono, el objetivo de este capítulo es desarrollar una aplicación que permita al tomador de decisiones analizar la información a partir de una base de datos multidimensional difusa. El desarrollo de esta aplicación se acentúa en dos planos conceptuales: en primer plano la aplicación de los algoritmos que permiten generar el valor difuso de las variables a difusificar como el elemento innovador de nuestra investigación y en un segundo plano la creación del modelo datos lógico y físico.

En la figura 4.10 se muestra el paquete de información utilizado para la generalización de cómo los datos difusos serán estructurados en el depósito de datos y en la figura 4.11 se ilustra el modelo lógico de datos derivado a partir del paquete de información.

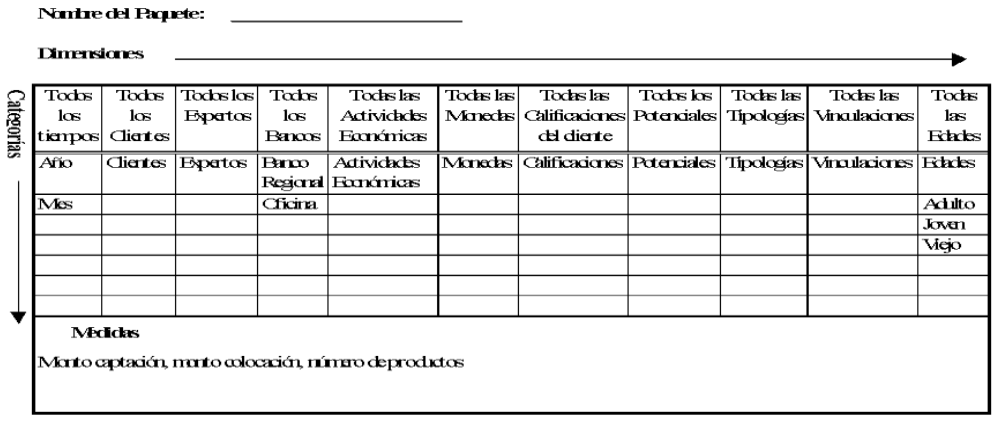


Figura 4.10 Paquete de información difuso

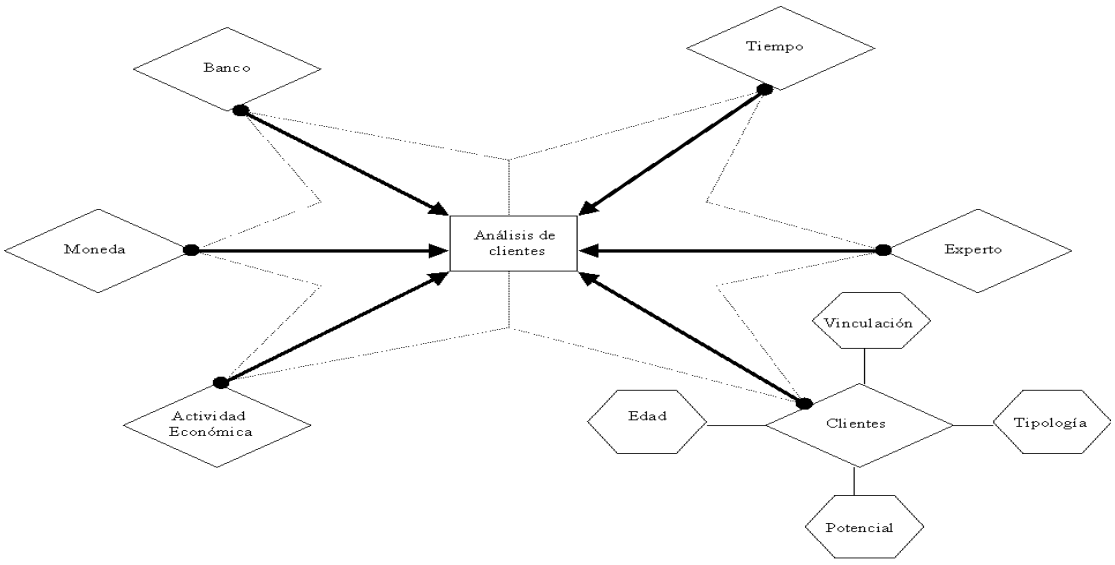


Figura 4.11 Modelo estrella difuso

Cada una de estas figuras es el resultado final de la aplicación de las tareas definidas en nuestro procedimiento. Finalmente, en la figura 4.12 se muestra el modelo físico de datos que da sustento al mercado de datos que incorpora la información difusa.

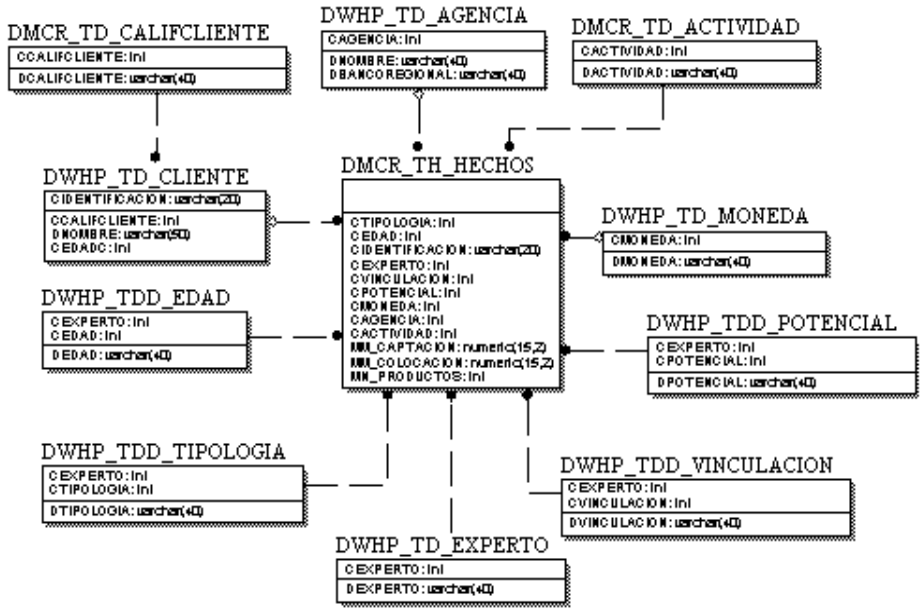


Figura 4.12 Modelo difuso físico

No se explica el detalle de las tareas realizadas para la generación del mercado de datos difuso, ya que estas actividades fueron explicadas con mayor propiedad en el capitulo 3 y en el desarrollo del caso práctico.

Todo lo anterior nos permite concluir que el modelo MVL y el procedimiento para la creación de un depósito de datos pueden ser combinados para creación de un depósito de datos difuso, herramienta a través de la cual el tomador de decisiones podrá responder a preguntas no puntuales, sino como una variable lingüística, por ejemplo, podrá analizar la cartera de clientes desde el punto de vista de su vinculación: ¿Cuáles clientes tienen vinculación baja, media o alta?. Desde el punto de vista de su potencial: ¿Cuáles clientes tienen potencial alto, medio/alto o bajo?. Cada una de estas preguntas serán evacuadas mediante la relación de las dimensiones y las medidas correspondientes, como se muestra en la siguiente figura.

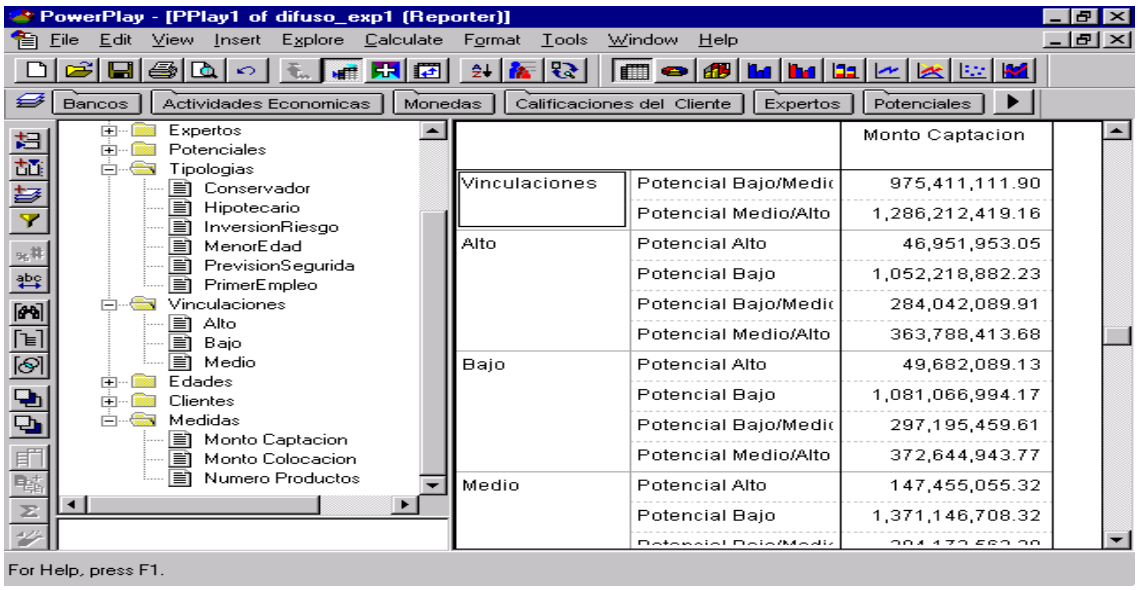


Figura 4.13 Base multidimensional difusa

En la figura 4.14 se muestra gráficamente un ejemplo del tipo de consultas que el usuario podrá realizar con esta herramienta, a fin de ampliar su panorama del negocio.

Nótese que el tomador de decisiones podrá relacionar de una manera muy fácil valores puntuales, por ejemplo, la calificación del cliente, con variables difusas: edad y potencial, lo que viene a enriquecer semánticamente la interpretación de los datos.

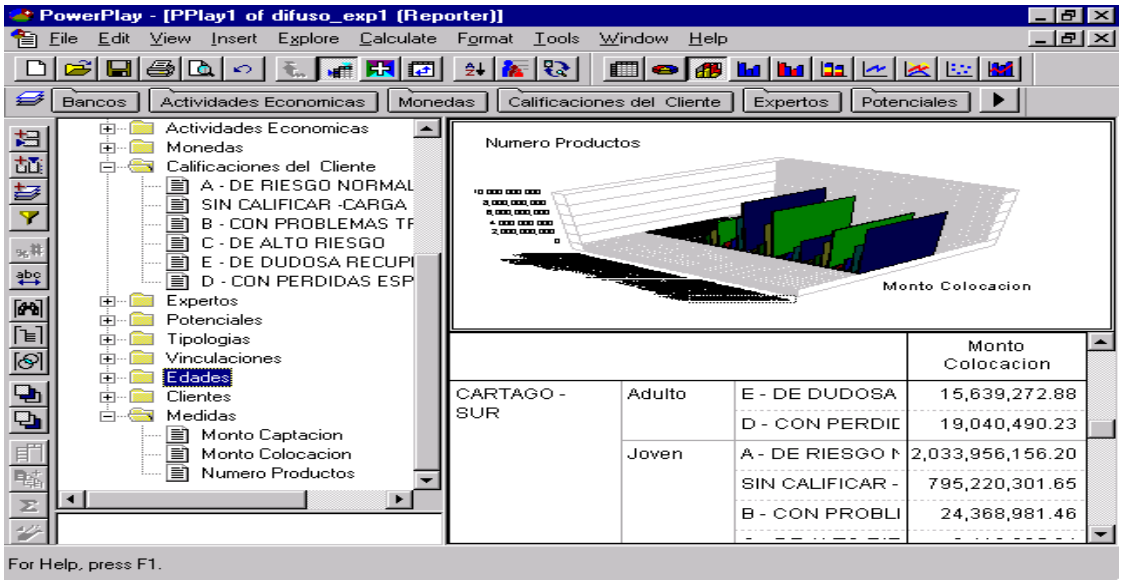


Figura 4.14 Consulta Difusa

De esta forma el tomador de decisiones visualiza la captación por región, y dentro de cada región clasificada por la edad del cliente, asimismo la edad del cliente es segmentada de acuerdo con la calificación de riesgo asignada a cada uno de los clientes.

Con esto se concluye la aplicación propuesta de este trabajo, el cual es el desarrollo de una metodología que permita construir un depósito de datos al cual se le pueda agregar información con características inciertas y que las mismas pueden ser modeladas mediante variables lingüísticas.

Entonces, este trabajo provee al usuario con un marco conceptual y una herramienta innovadora que le permite ser más eficiente en la toma de decisiones.

CAPÍTULO 5

CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

El acceso a la información es un elemento estratégico en la sociedad actual. Sin embargo, lo más importante no es la cantidad de información a la que se pueda acceder, si no la calidad de los mecanismos de que se disponen para acceder a aquella información que nos interesa en un momento determinado.

Dentro de este contexto los depósitos de datos permiten reunir en forma integrada, resumida y detallada los datos históricos de la organización, específicamente estructurados para la toma de decisiones. No obstante, a pesar de los esfuerzos realizados, la representación de la información y el tratamiento de la misma, se encuentra todavía lejos de las formas de expresión utilizadas habitualmente por el tomador de decisiones.

Tal inquietud nos permitió investigar todo lo relacionado con las herramientas para la toma de decisiones y las diferentes formas de representar la información imprecisa. Para alcanzar nuestro objetivo principal se elaboró un procedimiento que permitió la construcción de un depósito de datos que combine e integre aspectos que representen y manipulen información imprecisa por medio de variables lingüísticas.

Este trabajo ha enfocado de manera integral los problemas para mostrar y almacenar información imprecisa, creando para ello un procedimiento basado en las experiencias obtenidas del desarrollo de un depósito de datos y de la investigación de varias metodologías disponibles en el medio. Además, se crea el fundamento teórico para desarrollar el modelo MVL como mecanismo para solucionar los problemas de la representación y almacenamiento de datos imprecisos.

El empleo de una forma de trabajo ordenada es un factor de importancia en el desarrollo e implementación de un proyecto de depósito de datos, y la tendencia general

busca lograr, a través del uso de la metodología, recortar los tiempos de desarrollo y programar la inversión de recursos de manera eficiente.

En la actualidad no podemos asegurar cuál estrategia o metodología es la mejor, sin embargo, al analizar las tendencias generales del mercado, encontramos que las organizaciones están adoptando más frecuentemente alguna estrategia de desarrollo de depósitos de datos, la cual garantice la probabilidad de éxito en la implantación de este tipo de tecnología.

El procedimiento propuesto se fundamenta en la estrategia de plantear un proceso que consta de varias fases (arquitecturas) y que se pueden ajustar con cada nuevo requerimiento que se desee incluir en el depósito de datos, el resultado final del procedimiento es un nuevo producto que soportará las necesidades de los tomadores de decisiones, permitiendo de esta forma a las organizaciones que adopten nuestra metodología disminuir sustancialmente los tiempos de análisis y diseño de construcción de depósito de datos y asignar los recursos de manera eficiente.

Quizás el aporte más novedoso e importante de este procedimiento es la forma de cómo se conceptual izan y definen las variables que serán incluidas en la base de datos, a partir de los requerimientos de los usuarios. Además, de proveer al usuario el marco conceptual y una herramienta con la cual puede implementar su aplicación una vez llevado a cabo cada una de las tareas contempladas en el procedimiento propuesto.

En los aspectos prácticos, lo más importante es que los planteamientos teóricos se llevaron a aplicaciones concretas, de hecho la motivación para llevar a cabo esta investigación fue un problema real: construir un depósito de datos y la creación de una base de datos multidimensional con información de las colocaciones de una Institución Bancaria. Cabe destacar que la versatilidad de esta herramienta se explota ampliamente gracias a los aspectos desarrollados a través del procedimiento propuesto.

Asimismo, la aplicación sistemática de cada uno de los pasos descritos en cada fase es una innovación tecnológica importante, no solo porque abarata los costos y acelera el proceso de la toma de decisiones, sino también porque le da un valor agregado a la información utilizada para la toma de decisiones, que frecuentemente esta dispersa y se utiliza en forma aislada. Al integrar esta información en un depósito de datos que combine e integre aspectos que representen y manipulen información imprecisa, se puede hacer decisiones más adecuadas, mejorando la calidad y la oportunidad. También posibilita que la información pueda ser utilizada en otros tipos de inferencias y áreas del negocio.

Nótese que la facilidad que da nuestra metodología de desarrollar depósitos de datos con elementos difusos, da a esta investigación una amplia gama de aplicaciones, pues no solo lleva a cabo inferencias en el área financiera, sino que tiene otras aplicaciones potenciales en tanto se tenga el conocimiento necesario.

Es importante resaltar que queda bajo responsabilidad del usuario experto escribir las reglas que permiten obtener los valores asociados a cada una de las variables lingüísticas.

Entonces, este trabajo provee al usuario con un marco conceptual y una herramienta en la cual implementar su aplicación, una vez llevado a cabo el análisis y captura semántica del dominio del problema. Esto significa, que se establecen las clases, objetos, relaciones y restricciones que modelan la situación, con la ventaja de utilizar elementos basados en lógica difusa.

Por otro lado, esta investigación da nuevos aportes en aspectos teóricos, como es la representación de la información difusa mediante el empleo de dominios construidos sobre distribuciones de posibilidad. Con respecto a la manipulación de la información difusa, el modelo MVL presenta una gran flexibilidad para el tratamiento y evaluación de la información difusa, ya que por medio de las variables lingüísticas se estructura tanto la información de origen como la que resulte de las operaciones realizadas sobre los mismos.

Cabe destacar que la versatilidad del modelo MVL se debe esencialmente a la propuesta de almacenamiento que incluye una estructura auxiliar con formato flexible, que puede adaptarse a las necesidades del usuario, de manera que pueda almacenar un número indefinido de etiquetas lingüísticas asociada a un número difuso.

Por todo lo anterior, la originalidad de esta investigación se debe en gran parte a la combinación de valores discretos de un depósito de datos con valores difusos generados a partir de los criterios de los expertos, para generar como producto una base de datos multidimensional con valores difusos.

Como corolario de este esfuerzo de investigación se puede concluir que los sistemas de apoyo al proceso de tomas de decisiones se pueden enriquecer con el conocimiento y la información difusa capturada a partir del conocimiento y experiencia de los expertos de la organización en el área funcional sujeta a análisis. La figura 5.1 esquematiza a nivel general como se realiza este proceso.

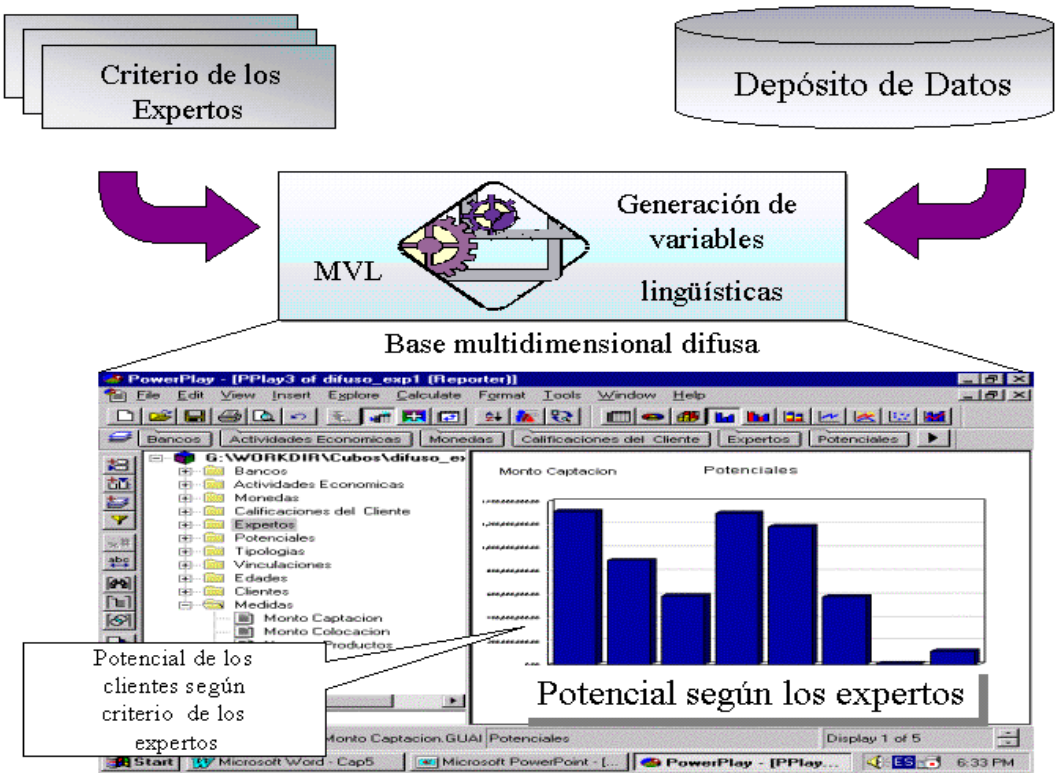


Figura 5.1 Esquema para la generación de una base multidimensional difusa, a partir del criterio de varios expertos.

Al incluir en este esquema el criterio de varios expertos, la novedad es que la información brindada por ellos puede ser aprovechada para otros fines, dando valor agregado a la información comercial tradicional, ampliando los criterios disponibles para la toma de decisiones, ya que se utiliza el criterio del experto de una forma sencilla y práctica.

Finalmente, la necesidad del análisis multidimensional oportuno, como soporte para la toma de decisiones, es cada vez más creciente. Como respuesta, los departamentos de sistemas se han inclinado por la tecnología de depósitos de datos para satisfacer la demanda de los usuarios. Sin embargo, es importante tomar en cuenta que existen diversas tecnologías orientadas a consultar, almacenar datos y analizar resultados.

5.2. RECOMENDACIONES

Esta investigación no pretende ser una solución total al problema de construcción de un depósito de datos que combine e integre aspectos que representen y manipulen información imprecisa, pero aborda aspectos relevantes para el almacenamiento y tratamiento de información difusa por medio de variable lingüísticas. Los resultados pueden ampliarse en varias direcciones de modo que capturen una semántica mayor de la lograda hasta ahora.

Un primer aspecto es extender el Modelo de Variables Lingüísticas (MVL) para que manipule otras formas de representar la información difusa, por ejemplo, implementar el concepto de relaciones difusas descrito por Raju & Majundar [RAJU1988], donde define las relaciones difusas basados en los conjuntos de difusos de manera análoga como la hace Codd, [CODD1970], fundamentada en su teoría de bases de datos, dado que es viable esta extensión mediante el empleo de los recursos que proporciona la teoría de conjuntos difusos.

Otra posible línea de investigación es la de incluir en el modelo la captura de mayor semántica con el ampliar la gama de consultas difusas. Para ello se sugiere la

implementación de la consulta difusa simple [KACP1986], con el cálculo difuso de predicados con su forma general:

Y es *F*

Dónde Y es una clase de objetos y F es alguna propiedad, lo cual permitiría hacer consultas como ¿Qué tan *buen*o es un *client*e?. Todo lo anterior implica, que basta con encontrar una forma de calcular el grado de pertenencia, o valor de verdad.

Otra área de extensión interesante de este trabajo, sería capacitar al tomador de decisiones el uso y representación de conceptos en términos difusos, para resolver problemas donde un objeto es mejor caracterizarlo con el cumplimiento mínimo de varios atributos que con el cumplimiento máximo de un solo atributo. En este sentido el usuario podría explotar mejor la información sujeta análisis, sin necesidad de tener un dominio total de los atributos del sistema de información.

La metodología propuesta y el modelo MVL constituyen un prototipo ideal para investigar el desarrollo de algoritmos eficientes para mejorar el rendimiento y masificar el producto en un proyecto de carácter institucional.

Es por ello que nos planteamos el por qué no realizar una definición formal de un modelo teórico lógico con capacidad para realizar deducción automática a partir de la información imprecisa e incierta. Este modelo ha de ser lo suficientemente flexible y general como para permitir integrar las características más relevantes de cada uno de los modelos de bases de datos que se encuentran en la literatura.

Además, es necesario estudiar y analizar una posible extensión del modelo MVL para la manipulación coherente y correcta de cualquier tipo de omisión de datos, esto es, información disyuntiva, desconocida, inaplicable o nula. Dotar a dicho modelo de un sistema de deducción, adecuado a los datos que se van a almacenar, esto conlleva la posibilidad de definir y utilizar reglas de carácter impreciso que serán aplicadas sobre los datos, y describir un mecanismo de deducción a partir de dichas reglas. Llevar a cabo la

construcción de un prototipo experimental que muestre las posibilidades de dicho modelo teórico.

Con respecto al desarrollo de depósitos de datos, en la actualidad no se puede asegurar cuál estrategia o metodología es la mejor o la peor, sin embargo, al analizar las tendencias generales del mercado, recomendamos que la estrategia de desarrollar mercados de datos para áreas muy específicas del negocio, es la que está siendo adoptada más frecuentemente en los últimos tiempos, a esta tendencia general se le ha identificado como la aproximación que garantiza la probabilidad de éxito más grande en la implementación de depósitos de datos.

Aunado a lo anterior, se conoce de pocos depósitos de datos exitosos, que satisfacen las demandas del usuario. En este sentido, los mercados de datos a base de OLAP, dentro o fuera de la arquitectura de depósitos de datos, ha demostrado ser una solución práctica para el usuario final.

BIBLIOGRAFÍA

[ALUR1995] Alur, N. "Missing Links in Data Warehousing". *Database Programming and Design*, 8(9), 21-23, Set. 1995

[AMST1996] Armstrong, R. "Data Warehousing; the fallacy of data mart centric strategies". NCR Corporation, 1996

[ARAB1997] Arabito, C. "Data Warehousing, The competitive advantage within reach". Tomado de Internet, dirección: www.sysameri.com/osc/wp32.htm, 1997

[BAKH1996] Bachman, R. Khabaza, T. "Mining Business Database". *Communication of ACM*, 39(11), 42-48, Noviembre 1996

[BAUM1996] Baum, D. "Data Warehouse; Building blocks for the next millennium". *Oracle Magazine*, x(2), 36-43, Marzo-Abril 1996

[BISC1994] Bischoff, J. "Achieving Warehouse success". *Database Programming and Design*, 27-33, Julio 1994

[BOAR1996] Board, B. "Understanding Data Warehousing Strategically". NCR Corporation, 1996

[CHAR1994] Charles, E. "The Paradoxical Success of Fussy Logic". *IEEE Expert*, 3-7, Agosto 1994

[CODD1970] Codd, E.F. "An Relational Model of Data for large shared data banks". *Communication of The ACM*, 13(6), 377-387, 1970

[CODD1986] Codd, E.F. "Missing Information (apllicable an inapplicable) in relational Data Base". *ACM Sigmod Record*, v. 15,4, 1986

[CODD1990] Codd, E.F. "*The Relational Model for Data Base Management*". Estados Unidos de América: Reading Addison Wesly, 1990

[COYO1990] Coad, P., Yourdon, E. "Object-Oriented Analysis". Yourdon Press, New Yersy, U.S.A., 1990

[DATE1990] Date, C.J. "An Introduction to Data Base System". Estados Unidos de América: Addison-Wesley Publishing Co., vol. 1.5 ed., 1990

[DELG1997] Delgado, R. “Bases de Datos Multidimensionales”. *Soporte Latino Oracle Latinoamérica*, v. 1, 32-35, 1997

[DEVL1997] Devlin, B. Data Warehouse from architecture to implementation. Addison-Wesley, 1997

[DUPR1992] Dubois, D.; Prade H. “Las lógicas de lo vago y de lo muy posible”. *Mundo Científico*. España, 12(120), 60-67, 1992

[FAPI1996] Fayyad, U. Piatetsky, G. “The KDD Process for Extracting Useful Knowledge from Volumnes of Data”. *Communication of the ACM*, 39(11), 27-34, Noviembre 1996

[FAUT1996] Fayyad, U. Uthurusamy, R. “Data Mining and Knowledge Discovery in Databases”. *Communications of the ACM*, 24-26, Noviembre 1996

[FRYE1996] Fryer, R. “Basements, Junkyards, and Warehouses”. NCR Corporation, Nov. 1996

[GAIN1983] Gaines, B. R. “Precise past – fuzzy future”. *International Journal of Man-Machine Studies*. U.K. v.19, 117-134, 1983

[GIRA1998] Gill, H., Rao, P. *Data Warehousing; La integración de información para la mejor toma de decisiones*. México: Prentice Hall Hispanoamericana, 1998

[GISM1996] Gismondi, N. “Data Warehousing”. Tomado de Internet, dirección: www.guia.com.uy/45/data45.htm, 1996

[GONZ1996] González, Carlos. “*Sistemas de Bases de Datos*”. Costa Rica: Editorial Tecnológica de Costa Rica, primera edición, 1996

[GUPT1997] Gupta, V. “An Introduction to Data Warehousing”. Tomado de Internet, dirección: www.system-services.com/dwintro.htm, Agosto 1997

[HAMM1996] Hammergren, T. *Data Warehousing; Building the Corporate Knowledge Base*. Estados Unidos de América: International Thomson Computer Press, 1996

[IMMA1996] Imielinski, T. Mannila, H. “A Database Perspective on Knowledge Discovery”. *Communication of ACM*, 39(11), 58-64, Noviembre 1996

[INGL1997] Inmon, W., Glassey. K. *Managing the Data Warehouse*. Estados Unidos de América: Wiley Computer Publishing, 1997

[INMO1995] Inmon, W. “What is a Data Warehouse”. Tomado de Internet, dirección: www.cait.wustl.edu/papers/prism/vol1_no1/subject, 1995

[INMO1996] Inmon, W. “The Data Warehouse and Data Mining”. *Communication of ACM*, 39(11), 49-50, Noviembre 1996

[INMO1999] Inmon, W. “What is a Data Mart”. D2K Incorporated, 1999

[KACP1986] Kacprzyk, J., Ziolkowski, A. “Database queries with fuzzy linguistic quantifiers”. *IEEE Transactions on Systems, Man, and Cibernetics*. (U.S.A). 16(3), 474-479. 1986

[KIMB1997] Kimball, R. “A Dimensional Modeling Manifesto”. *DBMS*, Agosto 1997

[KIMB1998] Kimball, R. The Data Warehouse; Lifecycle Toolkit. Estados Unidos de América: Wiley Computer Publishing, 1998

[KOSI1993] Korth, H., Silberschatz, A. “Fundamentos de Bases de Datos”. España: McGraw Hill, segunda edición, 1993

[KYRU1995] Kyszkiewicz, M., Rubinski, H.”Reducing Information System with uncertain attributes”. *ICS Research Report*, v.56, 1995

[LEWO1989] Leung, K., Wong, M. “ A fuzzy expert database system”. *Data & Knowledge Engineering*, v. 4, 287-304, 1989

[MADS1996] Madsen, M. “Warehouse Design in the Agrregate”. *Database Programming and Design*, 9(7), 45-51, Julio, 1996

[MADS1997] Madsen, M. “Warehousing meets the web”. *Database Programming and Design*, 38-43, Agosto 1997

[MENN1996] Menninger, D. “Designing a Database for OLAP”. *Oracle Magazine*, x(2), 83-87, Marzo-Abril 1996

[MORI1995] Moriarty, T. “A Data Warehouse Primer”. *Database Programming and Design*, 57-59, Julio 1995

[MORR1996a] Morris, H. “Vertical Warehouses; will they stand up as packaged solutions”. International Data Corporation, Junio 1996

[MORR1996b] Morris, H. “Ultimately, it’s a systems integration challenge”. International Data Corporation, Junio 1996

[MUND1997a] Mundy, J. “Build and Maintain a Data Mart Database”. *Database Web Advisor*, 15(5), 56-58, Mayo 1997

[MUND1997b] Mundy, J. “Build and Maintain a Data Mart Database”. *Database Web Advisor*, 56-58, Mayo 1997

[MURD1978] Murdick, R. “Sistemas de Información Basados en Computadoras para la Administración Moderna”. México: Editorial Diana, 1978

[PAGR1995] Pawlak, Z., Grzymala-Busse, J. “Rough Sets”. *Communication of the ACM*, 38(11), 1995

[PARS1995] Parsaye, K. “The Sandwich Paradigm – avoid the data dump”. *Database Programming and Design*, 8(4), 50-55, Abril 1995

[PARS1997] Parsaye, K. “OLAP and Data Mining; bridging the gap”. *Database Programming and Design*, 30-37, Feb. 1997

[PASL1986] Pawlak, Z., Slowinshi, K. “Rough clasification of patiens after highly selective vagotomy for duodenal ulcer”. *International Journal of Man-Machine Studies*, v. 24, 413-433, 1986

[PLAN1991] Plant, R. E.; Stone , N. D. “Knowledge-based systems in agriculture”. *Biological Resource Series.* McGraw-Hill, San Francisco, Estados Unidos de América, 364, 1991

[RAJU1988] Raju, K. V. S. V. N; Majundar, A. K. “Fussy functional dependencies and losslees join decomposition of fuzzy relational database systems”. *ACM Transactions on Database Systems*, 13(2), 129-166, 1988

[POE1995] Poe, V. “DataWarehouse; architecture is not infrastructure”. *Database Programming and Design*, 8(7), 25-31, Julio 1995

[RASI1992] Rasiowa, H. “Toward fuzzy logic”. *Fuzzy Logic for the Management of Uncertainty*. Eds. Zadeh, L. A.; Kacprzyk, J. Jhon Wiley & Sons, New York, Estados Unidos de Norteamérica, 121-139, 1992

[REES1997] Reese, J. “Making the DataWarehouse”. *Database Programming and Design*, 10-13, Sep. 1997

[ROBI1998] Robinson, T. “A front-line hope for a back-end tool”. Sentry Market Research, 1998

[SAAC1997] Saylor, M. Acharya. M. “True Relational OLAP”. *Database Journal*, Nov.- Dec 1995

[SABA1995] Saylor, M. Bansal, S. “Open Systems Decision Support”. *Data Management Review*, Enero 1995

[SIMO1995] Simon, A. “I want a Data Warehouse. So, What is it again?”. *Database Programming and Design*, 26-31, Diciembre 1995

[STEV1997] Stevens, A. “Oracle Data Mart Suite – A case Study”. Tomado de Internet, dirección: www.oracle.com/dm/designstudy.htm, 1997

[THOM1997] Thomsen, E. “Dimensional Hierarchies and formulas”. *Database Programming and Design*, 61-63, Junio 1997

[TURB1988] Turban, E. “Decision Support and Expert System”. Estados Unidos de América: Prentice Hall, 1988

[VAN1990] Van der Gaag, L. C. “Different notions of uncertainty in quasi probabilistic models” . *International Journal of Man-Machine Studies*, 33, 595-606, 1990

[WAND1996] Wand. Y. “Anchoring Data Quality Dimensions”. *Communication of ACM*, 39(11), 86-95, Noviembre 1996

[WHIP1997] Whipple, L. “OLAPping at the shores of analysis”. *Databas Advisor*, 48-53, Febrero 1997

[WHIT1995] White, C. “The Key to a Data Warehouse”. *Database Progr5amming and Design*, 8(2), 23-25, Feb. 1995

[WILL1998] Williams, G. “Decisión Support a Competitive Imperative”. Unite Fall Conference, Reno, Nevada, Nov. 1998

[YAWO1998] Yazdani, S., Wong, S. *Data Warehousing with Oracle*. Estados Unidos de América: Prentice Hall PTR, 1998

[YOUN1994] Youngworth, P. “Build a DataWarehouse Solution”. *Database Advisor*, 12(3), 48-51, Julio 1994

[KUYU1995] G. J. KULR y B. YUAN. “Fuzzy Sets and Fuzzy Logic: Theory and Applications”. Prentice Hall PTR, Upper Saddle River, NJ, 1995.

[ZADE1965] Zadeh, L.A. “Fuzzy Sets”. *Information and Control*, v.8, 338-353, 1965

APÉNDICE A
MODELO RELACIONAL

En este apéndice se presentan los conceptos generales del modelo relacional de bases de datos propuestos por E.F. Codd [CODD1970],) con el objetivo de analizar los aspectos más importantes del mismo.

Partiendo de la propuesta original descrita por Codd [CODD1970], [CODD1986], [CODD1990) para la definición del Modelo Relacional, se tiene que en principio un Sistema de Administración de Bases de Datos Relacionales (SABDR) debe cumplir dos características básicas:

- Los datos son percibidos por el usuario como tablas o relaciones.
- Los operadores de los que dispone el usuario, por ejemplo para las consultas, generan como resultado de su aplicación nuevas tablas a partir de las existentes.

Asimismo, los conceptos más relevantes y los elementos más representativos de los SABDR tienen que ver con las siguientes partes bien diferenciadas: la estructura, integridad y manipulación de los datos.

Estructura de los datos

La estructura de los datos viene caracterizada por el concepto de relación. Para precisar dicho concepto es necesario la introducción de algunos términos que también forman parte de la definición general del modelo relacional:

Un *dominio* es un conjunto, normalmente finito de todos los valores posibles que puede tomar un atributo. Una característica inicial de los valores del dominio es la atomicidad, en el sentido que exista una descomposición del mismo que aporte significado, sin embargo, actualmente un dominio puede llevar asociado un conjunto de operadores específicos del mismo. El producto cartesiano de una serie de dominios D_1, D_2, ..., D_n, que denotaremos por $D_1 \times D_2 \times ... \times D_n$ es el conjunto de todas las *tuplas* $(x_1,x_2,...,x_n)$ tales que para cualquier i, i = 1, ...,n se verifica que $x_i \in D_i$, entendiendo que

tupla es una de las filas o registros de la relación. Se llama *relación* o tabla al elemento principal del modelo, siendo cualquier subconjunto del producto cartesiano de dos o más dominios. Una *instancia* de la base de datos es un conjunto finito de relaciones finitas, es decir, aquella que contiene un número finito de tuplas. Se llama *cardinalidad* de una relación al número de tuplas que contiene. Se llama *grado de una relación* $R \subset D_1 \times D_2 \times ... \times D_n$, siendo n, el número de dominios que intervienen en su definición.

Normalmente una relación puede verse como una tabla de valores, en la que las columnas llevan asociado un nombre que llamaremos atributo. Los valores de un atributo asociado a la columna i, pertenecen al dominio D_i. Una relación R de atributos A_1, A_2, ..., A_n, define lo que se llama un *esquema de relación*, y lo denotaremos por $R(A_1, A_2, ..., A_n)$, de manera que una relación específica R_1, con un conjunto concreto de tuplas, se dice que una instancia o extensión de dicho esquema.

Se define *clave primaria* al conjunto de atributos que identifican unívocamente cada tupla de una relación, pudiendo existir varios de estos conjuntos para una relación dada, pero solamente se seleccionará uno de esto como *clave primaria*, quedando los restantes como *claves externas*.

Integridad de los datos

Una base de datos consiste en una configuración de datos que se supone representa una porción del mundo real. Ningún modelo de base de datos puede garantizar que esa representación corresponda con la realidad en todo momento, esto supondría, entre otras cosas, que la base de datos tendría que poseer un conocimiento, no solo sobre los datos, sino también sobre su significado.

Lo que sí puede y debe hacer un SABDR es no permitir que se introduzca información que no pueda ser identificada, ni que se haga referencia en un lugar de la base de datos a información que no exista en la misma. Estos son los enunciados

informales de las reglas de Identidad y Referencial respectivamente. En forma más precisa pueden ser formulados como sigue:

- *Regla de Identidad*. Ningún componente de la *clave primaria* de una relación base puede aceptar un valor nulo. Donde nulo significa que la información no se encuentra por alguna razón, por ejemplo la propiedad no sea aplicable o porque el valor sea desconocido.

- *Regla de Referencia*. La base de datos no puede contener valores para la *clave externa* que no hallen correspondencia con los adoptados por la clave primaria a que hacen referencia.

La *clave externa* es un atributo o conjunto de atributos de una relación, que para cada valor adoptado por los mismos, identifican unívocamente una tupla en otra relación. Esto implica que los atributos que componen la clave externa en la relación de partida han de tener correspondencia con los que conforman la clave primaria en la relación referenciada.

En el punto anterior hemos introducido la definición de *clave primaria*. La necesidad de que exista una *clave primaria* para cada relación estriba en que es la única forma de acceder de forma unívoca a cada tupla de la misma.

Manipulación de los datos

En este apartado abordaremos el problema de la manipulación de las estructuras que aparecen en un SABDR. Esta manipulación se hace por medio de algún lenguaje formal de manejo de datos [KOSI1993], [DATE1990). En este sentido distinguiremos entre el Álgebra Relacional y el Cálculo Relacional. La diferencia esencial entre ambos es que mientras que el álgebra proporciona una colección de operadores y operaciones explícitas, el cálculo proporciona una notación para definir la relación resultante de una petición dada.

a) *Álgebra Relacional*.

El *álgebra relacional* (AR) es un lenguaje de consulta procedural, ya que el usuario da instrucciones al sistema para ejecutar una serie de operaciones en la base de datos, que calcularán los resultados deseados.

El AR consta de un conjunto de operadores que toman como entrada una o dos relaciones para dar como salida una nueva relación. Los operadores básicos del álgebra relacional pueden agruparse de la siguiente manera:

- *Operación de asignación:* es una operación especial que asigna el resultado de otras operaciones, sobre relaciones a una nueva relación. Se incluye como operador para poder conservar los resultados.
- *Operaciones tradicionales sobre conjuntos:* son unión, intersección, diferencia y producto cartesiano. Para todas ellas, excepto el producto cartesiano, es necesario que las dos relaciones operando sean compatibles, es decir , que sean del mismo grado y que los i-ésimos atributos de las dos relaciones tengan el mismo dominio.
- *Operaciones especiales:* son selección, proyección reunión y división.

b) *Cálculo Relacional*.

El *cálculo relacional* es, al contrario que el álgebra relacional, un lenguaje no procedural de consulta, ya que el usuario sólo expresa aquella información que desea obtener de la consulta, pero sin especificar qué operaciones han de realizarse sobre la base de datos para obtenerla.

Existen dos versiones del cálculo relacional [KOSI1993], [DATE1990): El cálculo relacional de tuplas y el cálculo relacional de dominios, que se describen a continuación:

b.1 *Cálculo Relacional de Tuplas*

Una expresión en el cálculo relacional de tuplas, es una expresión de la forma:

$$\{P/P(t)\}$$

que representa el conjunto de todas las tuplas *t* que hacen verdadera la fórmula o predicado *P*. Usaremos *t[A]* para denotar el valor que tiene la tupla *t* para el atributo *A* y $t \in R$ para denotar que la tupla *t* está en la relación *R*. En una misma fórmula pueden aparecer varias variables tuplas. Se dice que una variable tupla es *libre* si no va cuantificada, ni universal, ni existencialmente.

Una formula del cálculo relacional de tuplas se compone de *átomos*. Un átomo tiene una de las siguientes formas:

- $s \in R$, donde s es una variable tupla y *R* es una relación.
- *s*[x] Θ *v*[y], donde *s* y *v* son variables tupla, *x* es un atributo sobre el que *s* está definida, *y* es un atributo sobre el que *v* está definida, y Θ es un operador de comparación (<,>,≥,≤,=). En particular, se requiere que los atributos *x* e *y* tengan dominios cuyos valores puedan ser comparados por medio de Θ.
- *s*[x] Θ *c*, donde *s*, *x* y Θ tienen el mismo significado que en el párrafo anterior, y *c* es una constante del dominio del atributo *x*.

Las fórmulas se construyen a partir de los átomos usando las siguientes reglas:

- Un átomo es una fórmula.
- Si P_1 es una fórmula, entonces también lo son (P_1) y $\neg(P_1)$.
- Si P_1 y P_2 son fórmulas, también los son $P_1 \wedge P_2$, $P_1 \vee P_2$ y $P_1 \Rightarrow P_2$.
- Si $P_1(s)$ es una fórmula que contiene una variable de tupla libre *s*, entonces $\exists s \in R(P_1(s))$ y $\forall s \in R(P_1(s))$ también son fórmulas.

b.2 *Cálculo Relacional de Dominios*

Esta segunda forma del cálculo relacional usa variables de dominio, que toman valores del dominio de un atributo, en vez de valores de una tupla completa.

Una expresión del cálculo relacional de dominios es de la forma:

$$\{<x_1,x_2,\ldots,x_n> / P(x_1, x_2, \ldots, x_n)\}$$

donde x_1, x_2,…, x_n representan variables de dominio y P representa una fórmula compuesta por átomos. Un átomo en el cálculo relacional de dominios tiene una de las siguientes formas:

- $<x_1,x_2,\ldots,x_n> \in R$, donde R es una relación de n atributos y las x_i son variables o constantes de dominio.
- $x \ \Theta \ y$, donde x e y son variables de dominio y Θ es un operador de comparación. Es requisito indispensable que los atributos x e y tengan dominios que puedan compararse por medio de Θ.
- $x \ \Theta \ c$, donde x y Θ se definen como en el caso anterior, y c es una constante del dominio del atributo sobre el que se mueve x.

Como puede verse, la representación de un SABDR, además de ser sencilla, permite identificar de forma muy natural el contexto de las interpretaciones.

EVOLUCIÓN DEL MODELO RELACIONAL

Hoy en día la información es uno de los factores que más peso específico tiene en el desarrollo de una organización. Por este motivo, cualquier institución que pretenda no quedar rezagada en su desarrollo debe estar al tanto de las técnicas que van surgiendo en el almacenamiento, transmisión y análisis de la información. Históricamente las bases de datos han sido las herramientas diseñadas para llevar a cabo las tareas de almacenamiento

y para proporcionar algunos de los mecanismos necesarios para el análisis de la información [GIRA1998].

El objetivo de una base de datos es el almacenar la información en forma conveniente, el permitir su modificación de forma segura y el de facilitar el proceso de recuperación de aquella información, que en un momento dado nos resulte necesaria, todo ello en un formato adecuado a nuestras necesidades [YAWO1998].

A lo largo de las últimas décadas han sido múltiples las aproximaciones surgidas para atender estos requisitos. Estos criterios a los que responden las diferentes clasificaciones que se pueden realizar sobre las mismas son: la organización de los datos y el tipo de los mismos. Con respecto a la organización de los datos, los principales enfoques apararecidos han sido el de redes, el jerárquico, en [DATE1990] puede encontrarse una descripción de los principales desarrollos construidos en torno a estos dos enfoques, el relacional [CODD1970] y últimamente, el orientado a objetos [COYO1990].

Es interesante hacer notar que, si bien la organización de los datos no tiene que estar relacionada con el tipo de datos que se pueden soportar, obviamente, ciertas organizaciones presentan mayor flexibilidad que otras, en cuanto a la diversidad de tipos que pueden tratar. No obstante, a pesar de los esfuerzos realizados, la representación de la información y el tratamiento de la misma, se encuentra todavía lejos de los mecanismos de expresión utilizados habitualmente por el ser humano. En este sentido, se está realizando un gran esfuerzo para resolver los problemas teóricos y prácticos relacionados con la elaboración de bases de datos más inteligentes.

La idea que subyace a este tipo de bases de datos, es la de facilitar mecanismos para almacenar y recuperar información siguiendo un esquema más próximo al empleado por el ser humano. Dentro de estos aspectos, se identifica la posibilidad de representar y manipular la información cuya semántica se encuentra más próxima al esquema humano de representación y la introducción de mecanismos que doten al sistema de capacidad para inferir información a partir de la que se encuentra almacenada en el sistema.

El primer aspecto conlleva la incorporación en las bases de datos de capacidad para representar y manipular información imprecisa. El segundo aspecto implica la integración de las bases de datos y un amplio abanico de disciplinas relacionadas con la Lógica.

Las propuestas que surgieron con relación al primer aspecto se distinguen de los sistemas de bases de datos tradicionales en que, en los primeros sistemas la información que se posee sobre un atributo, existe o no existe, no permitiéndose sobre su conocimiento ningún grado de incertidumbre, es decir no se pueden representar o tratar información del tipo "Juan es alto". Tampoco se contempla la obtención de la información en términos imprecisos a partir de aquella que es expresada en forma precisa, por ejemplo, de un atributo que almacene información sobre la edad de los clientes, aquellos individuos que sean "viejos".

En este sentido, las diferentes aproximaciones aparecidas en la literatura presentan su propia versión de cómo combinar los modelos de representación de conocimiento con la teoría de conjuntos difusos de Zadeh [ZADE1965]. La mayoría de los autores centran sus propuestas teóricas en una extensión del modelo relacional que contemple, en mayor o menor grado un tratamiento para este tipo de datos. Para mayor información se puede investigar los esfuerzos realizados en este sentido, tanto por su creador, (E. F. Codd), como por C. J. Date, que han dado como resultado la aparición de nuevas versiones del modelo.

APÉNDICE B
CONJUNTOS DIFUSOS

El concepto de Conjunto Difuso fue introducido por Zadeh [ZADE1965], motivado por su interés para el análisis de sistemas complejos de control. De forma más precisa podemos introducir la definición de conjunto difuso como sigue:

Definición 1. Un conjunto difuso A sobre un universo de discurso X es un conjunto de pares:

$$A = \{x, \mu_A(x) : x \in X, \mu_A(x) \in [0,1] \} \quad (1)$$

donde $\mu_A(x)$ se denomina grado de pertenencia de x a A.

Según esto, si la "edad" es un universo de discurso de "joven", el conjunto difuso que representa dicho concepto quedaría expresado en la forma:

$$\text{joven} = \{(20,1.00),\ldots(35,0.20)\}$$

El identificador "joven" con la connotación que lleva asociado un conjunto difuso recibe la denominación de "variable lingüística".

Existen varias notaciones para el concepto de conjunto difuso dependiendo de la naturaleza del universo de discurso sobre el que definamos un conjunto difuso. Algunas de las más importantes son:

- Dado un universo de discurso finito $X = \{x_1, x_2,\ldots, x_n,\}$, un conjunto difuso A se puede denotar como:

  $$A = \mu_1 / x_1 + \mu_2 / x_2 + \ldots + \mu_n / x_n \quad (2)$$

  donde μ_1 representa el grado de pertenencia de x_1, con i = 1,2,…,n. Generalmente los elementos con grado cero no se listan.

- Dado un universo de discurso infinito X, un conjunto difuso A sobre X se puede representar como:

  $$A = \int \mu_A (x_1) / x, \quad (3)$$

donde $\mu_A (x_1)$ es el grado de pertenencia de x.

Algunos conceptos sobre conjuntos difusos son:

- *Igualdad de conjuntos difusos*

  Definición 2. Dos conjuntos difusos A y B sobre X, se dicen iguales denotado como A = B, sii,

  $$\forall x \in X, \mu_A (x) = \mu_B (x) \qquad (4)$$

- *Igualdad de conjuntos en otro*

  Definición 3. Dados dos conjuntos difusos A y B sobre X, decimos que $A \subseteq B$, sii,

  $$\forall x \in X, \mu_A (x) \leq \mu_B (x) \qquad (5)$$

- *Soporte de un conjunto difuso*

  Definición 4. El soporte de un conjunto difuso A definido sobre X, es un subconjunto de dicho universo que satisface:

  $$\text{soporte (A)} = \{x \in X, \mu_A (x) > 0 \} \qquad (6)$$

- *Núcleo de un conjunto difuso*

  Definición 5. El núcleo de un conjunto difuso A definido sobre X, es un subconjunto de dicho universo que satisface:

  $$\text{núcleo (A)} = \{ x \in X, \mu_A (x) = 1 \} \qquad (7)$$

- *Altura de un conjunto difuso*

  Definición 6. La altura de un conjunto difuso A definido sobre X, se define como:

$$\text{Altura}(A) = \sup_{x \in X} \mu_A(x) = 1 \qquad (8)$$

- *Conjunto difuso normalizado*

  Definición 7. Un conjunto difuso A definido sobre X, se dice normalizado sii,

$$\exists x \in X,\ \mu_A(x) = 1 \qquad (9)$$

esta definición implica que Altura(A) = 1

OPERACIONES SOBRE CONJUNTOS DIFUSOS

Las principales operaciones sobre conjuntos difusos son la unión, la intersección y el complemento (ZADE1965], [CHAR1994], [LEWO1989).

- *Unión*

  Definición 8 Si A y B son dos conjuntos difusos sobre un universo de discurso X, la función de pertenencia de la unión de ambos conjuntos, $A \cup B$, viene dada por:

$$\mu_{A \cup B}(x) = f(\mu_A(x), \mu_B(x)),\ x \in X \qquad (10)$$

- *Intersección*

  Definición 9 Si A y B son dos conjuntos difusos sobre un universo de discurso X, la función de pertenencia de la intersección de ambos conjuntos, $A \cap B$, viene dada por:

$$\mu_{A \cap B}(x) = f(\mu_A(x), \mu_B(x)),\ x \in X \qquad (11)$$

- *Complemento*

  Definición 10. Una función C de [0,1] es el complemento si satisface las siguientes condiciones:

  - C = 1
  - C (C(a)) = a
  - C es estrictamente decreciente
  - C es continua

Aunque existen varios tipos de operadores que satisfacen tales propiedades, nosotros para el complemento, emplearemos principalmente la brindada por Zadeh, en la cual:

$$C(x) = 1- x \qquad (12)$$

Por tanto, para un conjunto difuso A sobre un universo de discurso X, la función de pertenencia del complemento de A, ¬ A, viene dada por:

$$\mu_{\neg A}(x) = 1 - (\mu_A(x), x \in X \qquad (13)$$

NÚMEROS DIFUSOS

El concepto de número difuso fue introducido por primera vez por Zadeh, con el propósito de analizar y manipular valores numéricos aproximados. El concepto ha sido refinado sucesivamente, entendiendo nosotros por número difuso lo siguiente:

Definición 11 Sea A un subconjunto difuso de $\Re$ y μ_A su función de pertenencia cumpliendo:

1. $\forall x, y \in \Re, \forall \mu_A (t) \geq \min \{ mu_A(x), \mu_A (y)\}$, es decir que es convexo.
2. μ_A es semicontinua superiormente.
3. El soporte de A es un conjunto acotado.

Entonces diremos que A es un número difuso.

Algunos autores incluyen en la definición la necesidad de que el subconjunto difuso esté normalizado.

La forma general de la función de pertenencia de un número difuso M, es la siguiente

$$\mu_M (x) = \begin{cases} r_M(x) & si \quad x \in [m-a, m] \\ \alpha_M \ si & x \in [m,n] \\ s_M(x) & si \quad x \in (n, n+b) \\ 0 & en\ otro\ caso \end{cases}$$

donde $r_{M,}\ s_{M} : \Re \rightarrow [0,1]$, r_M no decreciente, s_M no creciente, $r_M (m) = \alpha_{M =} s_M (n)$, $\alpha_M \in [0,1]$, y a, b, m, n $\in \Re$.

Al número α_M se le denomina *altura* del número difuso, al intervalo [n,m] *intervalo modal*, y a los números a y b holgura izquierda y derecha respectivamente. El número difuso de la figura B.1, es una representación de “aproximadamente entre m y n”.

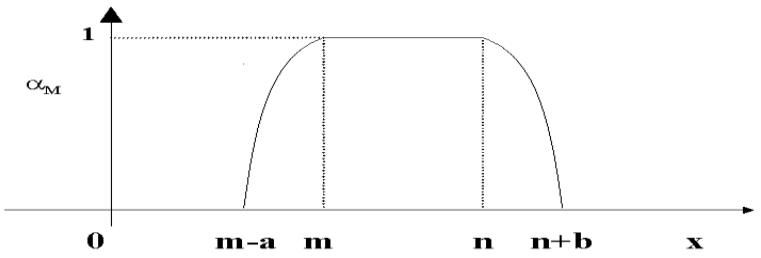


Figura B.1 Aspecto general de un número difuso

En nuestro trabajo, utilizaremos, un caso particular de número difuso, que se obtiene cuando consideramos a r_M y s_M como funciones lineales. En este caso la función de pertenencia adopta la forma:

$$\mu_M(x) = \begin{cases} \alpha_M + (x - m)\,\alpha_M / a & \text{si } x \in [m - a, m] \\ \alpha_M & \text{si } x \in [m, n] \\ \alpha_M - (x - n)\,\alpha_M / b & \text{si } x \in (n, n + b) \\ 0 & \text{en otro caso} \end{cases}$$

A un número difuso de este tipo lo llamaremos triangular o trapezoidal. Nosotros emplearemos los números difusos normalizados por lo que $\alpha_M = 1$, en este caso podremos caracterizar un número difuso trapezoidal normalizado M, mediante el empleo de los parámetros m,n,a,b como sigue:

$$M = (m,n,a,b)$$

En la figura B.2 se muestra una representación gráfica de dicho número.

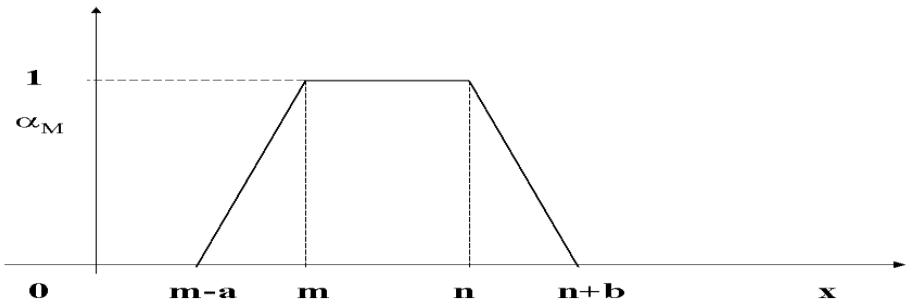


Figura B.2 Número difuso trapezoidal normalizado

De forma más general, un conjunto difuso A de un conjunto X en el sentido clásico puede definirse como un conjunto de pares ordenados cada uno con el primer elemento en X, y el segundo elemento en el intervalo [0,1], con exactamente un par ordenado presente para cada elemento de X. El valor de 0 se usa para representar una no pertenencia completa, el valor de 1 se usa para representar pertenencia completa. El conjunto X se refiere como el universo del discurso para el conjunto difuso A. Frecuentemente, el mapeo se describe como una función, la función de pertenencia de A.

El grado al cual la expresión x está en A es verdadera se determina al encontrar el par ordenado cuyo primer elemento es X.

Ejemplo 1, si se habla de las personas y su altura, el conjunto X es el conjunto de personas y se define un conjunto ALTOS, el cual responderá la pregunta en qué grado la persona X es alta?. Zadeh describe ALTO como una variable lingüística, la cual representa nuestra categoría cognoscitiva de ALTOS. Para cada persona en X, se tiene que asignar un grado de pertenencia en el conjunto difuso ALTOS, con la función de pertenencia basada en la altura de las personas:

$$\text{ALTO (X)} = \begin{cases} 0 & \text{si altura (x)} < 1.65 \text{ m.} \\ \dfrac{\text{altura (x)} - 1.65 \text{ m}}{0.66 \text{ m}} & \text{si } 1.65 \text{ m} \leq \text{altura (x)} \leq 2.31 \text{ m} \\ 1 & \text{si altura (x)} > 2.31 \text{ m} \end{cases}$$

El gráfico de esta función se muestra en la figura B.3

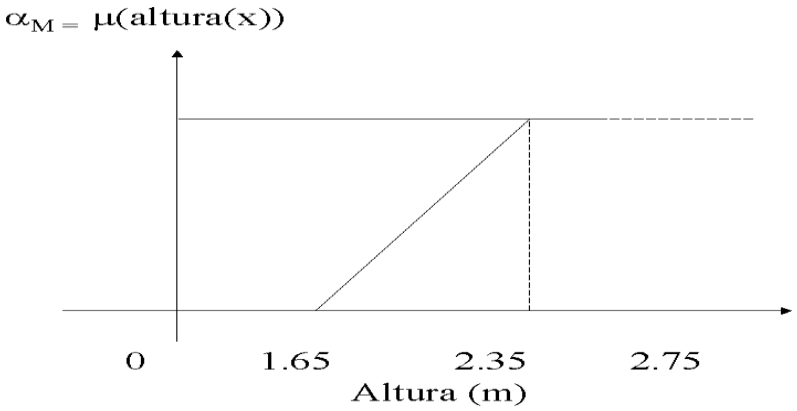


Figura B.3 Gráfico de la función de pertenencia del conjunto difuso ALTOS

APÉNDICE C

METADATO DE EXTRACCIÓN DE DATOS

SISTEMA: CREDITO

FUENTE DE INFORMACION	TRANSFORMACION	TABLA EN EL DWH
ARCHIVO TEXTO SISTEMA		**DWHP_TD_SISTEMA**
cte que identifica el sistema = 2		CSISTEMA
Sistema de Crédito (SICRE)		DSISTEMA
ARCHIVO TEXTO AGENCIAS_POR_BANCO (suministrado por crédito)		**DWHP_TD_AGENCIA**
cte que identifica el sistema = 2		CSISTEMA
NUMOFI		CAGENCIA
NOMBOFI		DNOMBRE
BANCO		DBANCOREGIONAL
SICRE.MG_EJECUTIVOS_DE_NEGOCIOS - SICRE.MG_FUNCIONARIOS – ARCHIVO EXTERNO GERENTES_POR_OFICINA (suministrado por crédito)		**DWHP_TD_EJECUTIVO**
cte que identifica el sistema = 2		CSISTEMA
	Se tomará ctipoejecutivo = 2 cuando la fuente son las tablas MG, ctipojecutivo = 1 para la fuente GERENTES_POR_OFICINA	CTIPOEJECUTIVO
mg_ejecutivos_de_negocio.codigo_ejecutivo or gerentes_por_oficina.codigo_agencia	Cuando se carga a partir de MG_EJECUTIVOS_DE_NEGOCIO.codigo_ejecutivo, GERENTES_POR_OFICINA.codigo_agencia	CEJECUTIVO
	asignación secuencial de valor	CSEJECUTIVO
mg_funcionario.codigo_agencia or gerentes_por_oficina.codigo_agencia	join entre MG_EJECUTIVOS_DE_NEGOCIOS con MG_FUNCIONARIOS, mediante el codigo_ejecutivo para tomar de la tabla MG_FUNCIONARIOS.codigo_agencia para la fuente MG. Cuando la fuente es el archivo txt se toma gerentes_por_oficina.codigo_agencia	CAGENCIA

SISTEMA: CREDITO

mg_ejecutivos_de_negocios.nombre or gerentes_por_oficina.nombre	Siempre que la fuente sea un archivo txt el nombre se toma de gerentes_por_oficina.GERENTE(nombre del ejecutivo), cuando las fuentes son las tablas MG se toma MG_ejecutivos_de_negocios.nombre	DNOMBRE
ARCHIVO TEXTO GIE (Suministrado por Dir. de Crédito)		**DWHP_TD_GRUPOINTERES**
cte que identifica el sistema = 2		CSISTEMA
num_grupo		CGRUPOINTERES
nom_grupo		DGRUPOINTERES
SICRE.PR_CALIFICACION_CARTERA		**DMCR_TD_CALIFCLIENTE**
codigo_calificacion		CCALIFCLIENTE
descripción		DCALIFCLIENTE
SICRE.MG_MONEDAS		**DWHP_TD_MONEDA**
cte que identifica el sistema = 2		CSISTEMA
codigo_moneda		CMONEDA
descripción		DMONEDA
SICRE.MG_ORIGEN_RECURSOS		**DMCR_TD_FONDEO**
cte que identifica el sistema = 2		CSISTEMA
codigo_origen		CFONDEO
clase_fondos	Si clase_fondos = 1 es Propio, si es 2 entonces Externo	DTIPOFONDEO
descripcion		DFONDEO
SICRE.MG_ACTIVIDADES_ECONOMICAS		**DMCR_TD_ACTIVIDAD**
cte que identifica el sistema = 2		CSISTEMA
codigo_actividad_economica		CACTIVIDAD
descripcion		DACTIVIDAD

SISTEMA: CREDITO

SICRE.MG_RUBRO_ACTIVIDAD **SICRE.MG_TIPO_INVERSION**		**DMCR_TD_RUBRO**
cte que identifica el sistema = 2		CSISTEMA
mg_rubro_actividad.codigo_actividad_economica		CACTIVIDAD
mg_rubro_actividad.codigo_inversion		CRUBRO
mg_tipo_inversion.descripcion	select a.codigo_actividad_economica, a.codigo_inversion, b.descripcion from mg_rubro_actividad a,mg_tipo_inversion b where a. CODIGO_INVERSION = b.CODIGO_INVERSION	DRUBRO
	Queda definida para posterior poblamiento de datos	**DMCR_TD_CLASE**
		CSISTEMA
		CACTIVIDAD
		CRUBRO
		CCLASE
		DCLASE
SICRE.PR_ESTADOS_CONTABLES		**DMCR_TD_ESTADOMORA**
cte que identifica el sistema = 2		CSISTEMA
codigo_estado_contable		CESTADOMORA
descripcion		DESTADOMORA
SICRE.MG_SUB_APLICACIONES		**DMCR_TD_TIPOCARTERA**
cte que identifica el sistema = 2		CSISTEMA
codigo_sub_aplicacion	Siempre que codigo_aplicación = "BPR"	CTIPOCARTERA
descripcion		DTIPOCARTERA

SISTEMA: CREDITO

SICRE.MG_CLIENTES - ARCHIVO EXTERNO MIGRUPOS (suministrado por la Dir de Crédito)		**DWHP_TD_CLIENTE**
cte que identifica el sistema = 2		CSISTEMA
codigo_cliente		CCODCLIENTE
numero_identificacion		CIDENTIFICACION
		CCUC
	Se toma DWHP_TD_CLIENTE.cccodcliente y se busca en pr_prestamos cada una de las operaciones activas del cliente (pr_prestamos.estado = 1),si la operación es normal, cobro_judicial o inactiva el saldo (pr_saldo_prestamos.valor) seria los registro cuyo pr_saldos_prestamo.codigo_tipo_saldo =1, si la operación esta en reserva de prestamo entonces pr_saldos_prestamo.codigo_tipo_saldo = 65. De cada una de sus operaciones se debe obtener sus respectivos saldos, para lograrlo se realizarra un join entre pr_prestamos y pr_saldos_prestamo donde los campos de union son pr_prestamos.(codigo_empresa,codigo_agencia,codigo_sub_aplicacion,numero_prestamo) en pr_saldos_prestamo.(codigo_empresa,codigo_agencia, codigo_sub_aplicacion,numero_prestamo) , sin tomar encuenta los prestamos cuya pr_prestamos.actividad_economica = 9, la asignación del criterio de endeudamiento será: 21 si la suma de pr_saldos_prestamos.valor de cada operacion es mayor a 5 Millones, 22 si la suma de pr_saldos_prestamos.valor de cada operación es <= 5 Millones	CNIVELENDEUDAMIENTO
codigo_categoria		CCALIFCLIENTE
codigo_ejecutivo	Si no tiene código de ejecutivo, se toma el mg.clientes.codigo_agencia del cliente, si tiene ejecutivo se asigna mg_clientes.codigo_ejecutivo. En caso que el cliente tenga operaciones en mas de una oficina, la oficina que definirá el código de ejecutivo será aquella que tenga el mayor endeudamiento, el cual se localiza en PR_SALDOS_PRESTAMOS, para los pr_saldos_prestamos.codigos_tipo_saldo = 1	CEJECUTIVO

SISTEMA: CREDITO

migrupos.num_grupo	El grupo económico asociado al cliente deberá ser buscado(codigo_identificacion) en el la tabla MIGRUPOS el cual es un archivo externo suministrado por crédito, en caso que el codigo_identificacion no exista en esta tabla, se le asignara un valor de cgrupointeres = 0	CGRUPOINTERES
	Si MG_CLIENTES.tipo_de _persona es F or N entonces "Fisica", si J entonces "Jurídica"	DTIPOCLIENTE
	Si MG_CLIENTES.tipo_de _persona es J, el nombre del cliente viene en la MG_CLIENTES.razon_social, en caso contrario se concatena MG_CLIENTES.nombres mas MG_CLIENTES.primer_apellido mas MG_CLIENTES.segundo_apellido	DNOMBRE
	SI MG_CLIENTES.codigo_ejecutivo no tiene valor asignado entonces cejecutivo = 1, en caso contrario =2	CTIPOEJECUTIVO
	Se programara	CSEJECUTIVO
	Queda definida para posterior población de datos	**DMCR_TD_SUBCLASE**
		CSISTEMA
		CACTIVIDAD
		CRUBRO
		CCLASE
		CSUBCLASE
		DSUBCLASE
ARCHIVO EXTERNO TPRODUCT		DMCR_TD_TIPOPRODUCTO
cte que identifica el sistema = 2		CSISTEMA
tproduct.tipo_producto		CTIPOPRODUCTO
tproduct.nombre		DTIPOPRODUCTO

SISTEMA: CREDITO

ARCHIVO EXTERNO PRODUCTO		**DMCR_TD_PRODUCTO**
cte que identifica el sistema = 2		CSISTEMA
producto.tipo_producto		CTIPOPRODUCTO
producto.producto		CPRODUCTO
producto.nombre		DPRODUCTO
ARCHIVO EXTERNO SPRODUCT		**DMCR_TD_SUBPRODUCTO**
cte que identifica el sistema = 2		CSISTEMA
sproduct.tipo_producto		CTIPOPRODUCTO
sproduct.producto		CPRODUCTO
sproduct.subproducto		CSUBPRODUCTO
sproduct.nombre		DSUBPRODUCTO
	Se parametriza para futura poblacion	**DMCR_TD_CALIFCREDITO**
		CSISTEMA
		CCALIFCREDITO
		DCALIFCREDITO
SICRE.GA_TIPO_GARANTIAS		**DWHP_TD_TIPOGARANTIA**
cte que identifica el sistema = 2		CSISTEMA
codigo_tp_garantia		CTIPOGARANTIA
descripcion		DTIPOGARANTIA
SICRE.MG_APROBACION		DMCR_TD_NIVELPROBACION
cte que identifica el sistema = 2		CSISTEMA
cte=1		CPERIODO
codigo_aprobacion		CNIVELAPROBACION
descripcion		DNIVELAPROBACION

SISTEMA: CREDITO

		Nota ademas se debe crear un registro con la siguientes características, 2,0,0,"Antes de 1998"
ARCHIVO EXTERNO COMPCART		**DMCR_TD_COMPCARTERA**
cte que identifica el sistema = 2		CSISTEMA
compcart.codigo_compcartera	1= Activa, 2= Inactiva	CCOMPCARTERA
compcart.nombre		DCOMPCARTERA
ARCHIVO EXTERNO SECCION		**DMCR_TD_SECCION**
cte que identifica el sistema = 2		CSISTEMA
seccion.codigo_seccion	Si es 1 = Corto Plazo, 2= Mediano, 3= Largo	CSECCION
seccion.nombre		DSECCION
SICRE.PR_SECCIONES		**DMCR_TD_SUBSECCION**
cte que identifica el sistema = 2		CSISTEMA
codigo_actividad_economica		CACTIVIDAD
	se carga tantas veces exista secciones	CSECCION
codigo_seccion		CSUBSECCION
descripcion		DSUBSECCION
SICRE.PR_PRESTAMOS - SICRE.PR_SALDOS_PRESTAMO - ARCHIVO EXTERNO FECORTE - SICRE.MG_SUB_APLICACIONES - SICRE.CJ_COBROS_JUDICIALES		**DMCR_TH_CREDITO**
cte que identifica el sistema = 2		CSISTEMA
PR_PRESTAMOS.CODIGO_CLIENTE		CCODCLIENTE

SISTEMA: CREDITO

	CON CCODCLIENTE y CSISTEMA SE VA ABUSCAR EN LA TABLA DWHP_TD_CLIENTE obteniendo el campo dwhp_td_cliente.cidentificacion	CIDENTIFICACION
fecorte.fecha_de_corte de la información de crédito		CFSALDO
pr_prestamos.fecha_apertura		CFFORMALIZACION
pr_prestamos.fecha_vencimiento		CFVENCIMIENTO
pr_prestamos.fecha_pago_interes or pr_prestamos.Fecha_pago_interes_anticipa	Si pr_prestamos.interes_vencidos = 'S' se toma pr_prestamos.fecha_pago_interes. Si pr_prestamos.interes_vencidos = 'N' se toma pr_prestamos.fecha_pago_interes_anticipa	CFSERVICIOINTERES
pr_prestamos.Fecha_debe_desde		CFAMORTIZACION
pr_prestamos.Fecha_Documento		CFDOCUMENTO
pr_prestamos.numero_prestamo		COPERACION
pr_prestamos.codigo_sub_aplicación		CTIPOCARTERA
mg_sub_aplicaciones.codigo_moneda	Con la constante mg_sub_aplicaciones.codigo_aplicación = 'BPR' más pr_prestamos.codigo_sub_aplicación ir a mg_sub_aplicaciones y tomar el campo codigo_moneda	CMONEDA
pr_prestamos.codigo_origen		CFONDEO
pr_prestamos.estado_contable_act		CESTADOMORA
pr_prestamos.codigo_Actividad_Economica		CACTIVIDAD
pr_prestamos.codigo_inversion		CRUBRO
	No se poblara	CCLASE
	No se poblara	CSUBCLASE
pr_prestamos.identificacion_plazo		CSECCION
pr_prestamos.codigo_seccion		CSUBSECCION
	Si pr_prestamos.causacion_suspendidad = S entonces ccompcartera = 2, caso contrario ccompcartera = 1	CCOMPCARTERA

SISTEMA: CREDITO

	Si la pr_prestamos.fecha_apertura < = 31/12/98 entonces cnivelaprobacion = 0, caso contrario cnivelaprobacion = pr_prestamos.codigo_aprobacion	CNIVELAPROBACION
ga_garantias.codigo_tp_garantia. Este hecho se deja previsto para futuro poblamiento.	Para obtener el tipo de garantía (ga_garantias.codigo_tp_garantia) se realiza un join entre pr_prestamos y pr_prestamo_garantias, donde los campos de union son pr_prestamos(codigo_empresa,codigo_agencia,codigo_sub_aplicación,numero_contrato) join con pr_prestamo_garantias(codigo_empresa,codigo_agencia,codigo_sub_aplicación,numero_contrato) para obtener el pr_prestamo_garantias.numero_garantia, este pr_prestamo_garantias.numero_garantia join con ga_garantias.numero_garantia para localizar el ga_garantias.codigo_tp_garantia	CTIPOGARANTIA
Este hecho no se poblara dado que la categoría del riesgo se da en términos del cliente y no a nivel de cada operación	No se poblara	CCALIFCREDITO
	(cte =1)	CTIPOPRODUCTO
	(cte =01)	CPRODUCTO
	Parametrizado para futuras poblaciones	CSUBPRODUCTO
pr_prestamos.codigo_agencia		CAGENCIA
	Si pr_prestamos.fecha_apertura <= 31/12/1998 entonces cperiodo = 0 sino =1	CPERIODO
pr_prestamos.monto_inicial		MMONTOINICIAL

SISTEMA: CREDITO

pr_saldos_prestamo.valor	Si la operacion es normal, cobro_judicial o inactiva (pr_prestamos.codigo_estado_cartera= A, B o C respectiva) el saldo del capital será el valor contenido en pr_saldos_prestamo.valor cuyo pr_saldos_prestamos. codigo_tipo_saldo =1, si la operacion esta en reserva(pr_prestamos.codigo_estado_cartera= D) entonces el saldo del capital será el valor contenido en pr_saldos_prestamo.valor cuyo pr_saldos_prestamo.codigo_tipo_saldo = 65. Para obtener el saldo del capital, se realizara un join entre pr_prestamos y pr_saldos_prestamo donde los campos de union son pr_prestamos.(codigo_empresa,codigo_agencia,codigo_sub_aplicacion,numero_prestamo) en pr_saldos_prestamo.(codigo_empresa,codigo_agencia,codigo_sub_aplicacion,numero_prestamo)	MSALDOCAPITAL
pr_saldos_prestamos	Si la operacion es normal, cobro_judicial o inactiva (pr_prestamos.codigo_estado_cartera= A, B o C respectiva) el saldo de interés será el valor contenido en pr_saldos_prestamo.valor cuyo pr_saldos_prestamos .codigo_tipo_saldo =2 and 10, si la operacion esta en reserva(pr_prestamos.codigo_estado_cartera= D) entonces el saldo de interés será el valor contenido en pr_saldos_prestamo.valor cuyo pr_saldos_prestamo.codigo_tipo_saldo = 46 y 63 65 SUM (pr_saldos_prestamos.valor) Con el pr_prestamos.codigo_empresa,pr_prestamos.agencia,pr_prestamos.codigo_sub_aplicacion,pr_prestamos.numero_prestamo,con pr_saldos_prestamo.codigo_empresa,pr_saldos_prestamo.codigo_agencia,pr_saldos_prestamo.codigo_sub_aplicacion,pr_saldos_prestamo.numero_prestamo	MSALDOINTERES
	Suma de DMCR_TH_CREDITO.MSALDOCAPITAL más DMCR_TH_CREDITO.MSALDOINTERES	MSALDOTOTAL
	No se tiene tomara valor = 0	MMULTAATRASO

SISTEMA: CREDITO

cj_cobros_judiciales.valor_cobro_jucio	SUM(cj_cobros_judiciales.valor_cobro_juicio) Con el pr_prestamos.codigo_empresa,pr_prestamos.agencia,pr_prestamos.codigo_sub_aplicacion,pr_prestamos.numero_prestamo realiza un join cj_cobros_judiciales.codigo_empresa,cj_cobros_judiciales.agencia,cj_cobros_judiciales.codigo_sub_aplicacion,cj_cobros_judiciales.numero_cuenta y se toma el cj_cobros_judiciales.valor_cobro_juicio, para todos aquellos registros en que cj_cobros_judiciales.situacion = A	MGASTOSJUDICIALES
cj_cobros_judiciales.valor_abogado	SUM(cj_cobros_judiciales.valor_abogado) Con el pr_prestamos.codigo_empresa,pr_prestamos.agencia,pr_prestamos.codigo_sub_aplicacion,pr_prestamos.numero_prestamo realiza un join cj_cobros_judiciales.codigo_empresa,cj_cobros_judiciales.agencia,cj_cobros_judiciales.codigo_sub_aplicacion,cj_cobros_judiciales.numero_cuenta y se cj_cobros_judiciales.valor_abogado, para todos aquellos registros en que cj_cobros_judiciales.situacion = A	MHONORARIOS
	SUM(pr_seguros.cobertura_poliza) Con el pr_prestamos.codigo_empresa,pr_prestamos.agencia,pr_prestamos.codigo_sub_aplicacion,pr_prestamos.numero_prestamo realiza un join pr_seguros.codigo_empresa,pr_seguros.agencia,pr_seguros.codigo_sub_aplicacion,pr_seguros.numero_prestamo y se toma pr_seguros.cobertura_poliza, siempre que pr_seguros.fecha_final => a la fecha de corte	MGASTOSPOLIZA
	Con el pr_prestamos.codigo_empresa,pr_prestamos.agencia,pr_prestamos.codigo_sub_aplicacion,pr_prestamos.numero_prestamo,se localiza en pr_saldos_prestamo.valor para el pr_saldos_prestamo.codigo_empresa,pr_saldos_prestamo.agencia,pr_saldos_prestamo.codigo_sub_aplicacion,pr_saldos_prestamo.numero_prestamo,pr_saldos_prestamos.codigo_tipo_saldo = 35(comisiones)	MCOMISION
	No se tiene tomara valor = 0	MGASTOSESTUDIOS
pr_prestamos.tasa_total		MTASAINTERES
pr_prestamos.plazo		MPLAZO
	No se poblara	MARREGLOS
	No se poblara	MINCUMPLIARREGLO

SISTEMA: CREDITO

pr_prestamos.cantidad_cuotas_pagadas		MNUMCUOTASCONT
pr_prestamos.numero_cuotas		MNUMCUOTASPAGAR
pr_prestamos.dias_atraso_intereses		MDIASATRASOINTERES
	Utilizar procedimiento empaquetado del Sistema de Crédito, él cual permite obtener los días de atraso.	MDIASATRASOCAPITAL
SICRE.PR_PRESTAMOS - SICRE.PR_SALDOS_PRESTAMO - ARCHIVO EXTERNO FECORTE - SICRE.MG_SUB_APLICACIONES - SICRE.CJ_COBROS_JUDICIALES		**DMCR_TH_CUOTAS**
cte que identifica el sistema = 2		CSISTEMA
pr_prestamos.codigo_agencia		CAGENCIA
PR_PRESTAMOS.CODIGO_CLIENTE		CCODCLIENTE
	CON CCODCLIENTE y CSISTEMA SE VA ABUSCAR EN LA TABLA DWHP_TD_CLIENTE obteniendo el campo dwhp_td_cliente.cidentificacion	CIDENTIFICACION
fecorte.fecha_de_corte de la información de crédito		CFSALDO
pr_prestamos.fecha_apertura		CFFORMALIZACION
pr_prestamos.fecha_vencimiento		CFVENCIMIENTO
pr_prestamos.fecha_pago_interes or pr_prestamos.Fecha_pago_interes_anticipa	Si pr_prestamos.interes_vencidos = 'S' se toma pr_prestamos.Fecha_pago_interes. Si pr_prestamos.interes_vencidos = 'N' se toma pr_prestamos.Fecha_pago_interes_anticipa	CFSERVICIOINTERES
pr_prestamos.Fecha_debe_desde		CFAMORTIZACION
pr_prestamos.Fecha_Documento		CFDOCUMENTO
pr_prestamos.numero_prestamo		COPERACION
pr_prestamos.codigo_sub_aplicación		CTIPOCARTERA

SISTEMA: CREDITO

mg_sub_aplicaciones.codigo_moneda	Con la constante mg_sub_aplicaciones.codigo_aplicación = 'BPR' más pr_prestamos.codigo_sub_aplicación ir a mg_sub_aplicaciones y tomar el campo codigo_moneda	CMONEDA
pr_prestamos.codigo_origen		CFONDEO
pr_prestamos.estado_contable_act		CESTADOMORA
pr_prestamos.codigo_Actividad_Economica		CACTIVIDAD
pr_prestamos.codigo_inversion		CRUBRO
	No se poblara	CCLASE
	No se poblara	CSUBCLASE
pr_prestamos.identificacion_plazo		CSECCION
pr_prestamos.codigo_seccion		CSUBSECCION
	Si pr_prestamos.causacion_suspendidad = S entonces ccompcartera = 2, caso contrario ccompcartera = 1	CCOMPCARTERA
	Si la pr_prestamos.fecha_apertura < = 31/12/98 entonces cnivelaprobacion = 0, caso contrario cnivelaprobacion = pr_prestamos.codigo_aprobacion	CNIVELAPROBACION
ga_garantias.codigo_tp_garantia. Este hecho se deja previsto para futuro poblamiento.	Para obtener el tipo de garantia (ga_garantias.codigo_tp_garantia) se realiza un join entre pr_prestamos y pr_prestamo_garantias, donde los campos de union son pr_prestamos(codigo_empresa,codigo_agencia,codigo_sub_aplicación,numero_contrato) join con pr_prestamo_garantias(codigo_empresa,codigo_agencia,codigo_sub_aplicación,numero_contrato) para obtener el pr_prestamo_garantias.numero_garantia, este pr_prestamo_garantias.numero_garantia join con ga_garantias.numero_garantia para localizar el ga_garantias.codigo_tp_garantia	CTIPOGARANTIA

SISTEMA: CREDITO

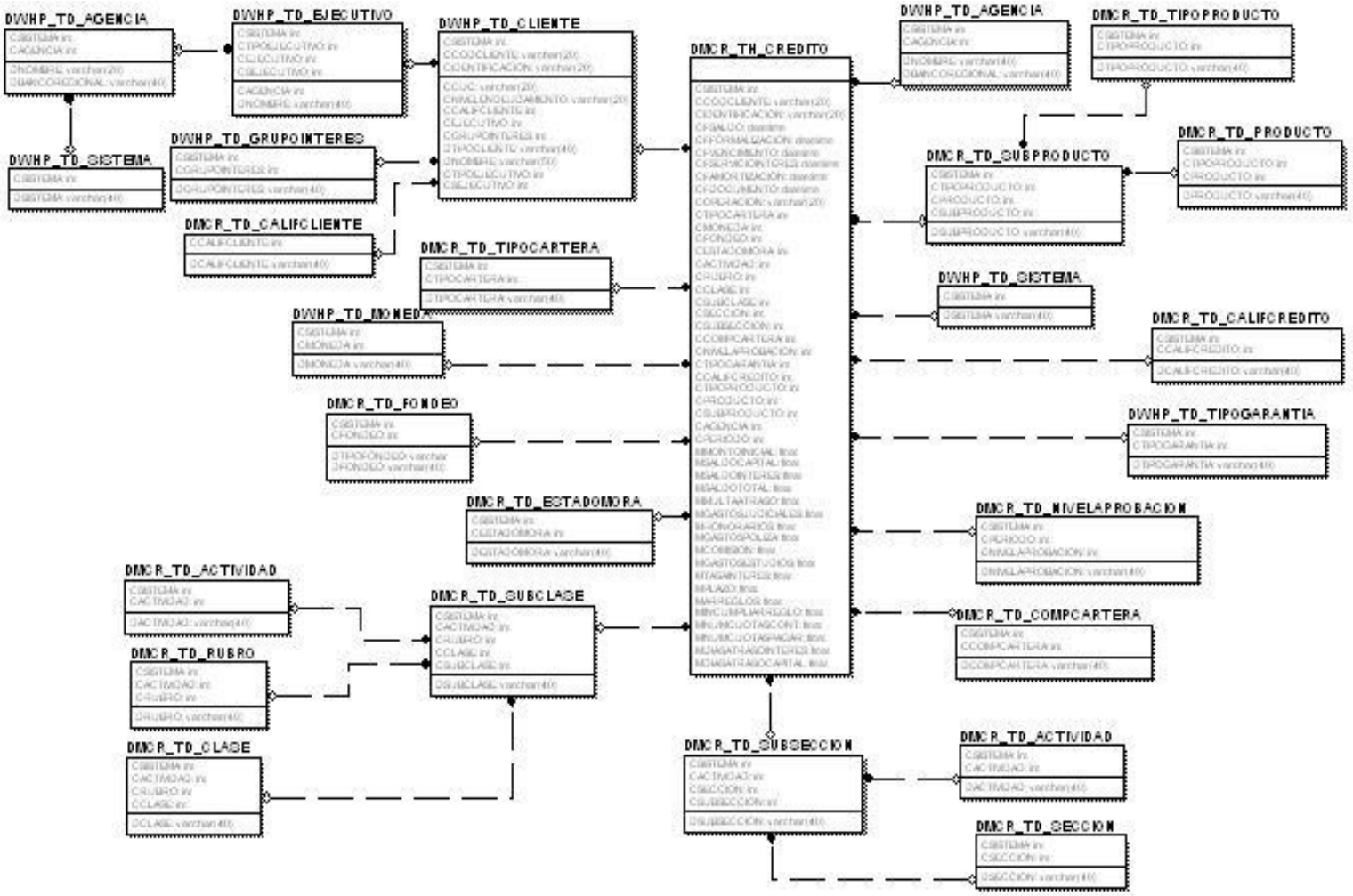
Este hecho no se poblara dado que la categoría del riesgo se da en términos del cliente y no a nivel de cada operación	No se poblara	CCALIFCREDITO
	(cte =1)	CTIPOPRODUCTO
	(cte =01)	CPRODUCTO
	Parametrizado para futuras poblaciones	CSUBPRODUCTO
	Si pr_prestamos.fecha_apertura <= 31/12/1998 entonces cperiodo = 0 sino =1	CPERIODO
	Parametrizado para futuras poblaciones	MSALDOAJUSTE
	El saldo de cuota será la suma de pr_saldos_prestamo.valor excepto pr_saldos_prestamos .codigo_tipo_saldo 3 y 1 Para obtener el saldo de cuota, se realizara un join entre pr_prestamos y pr_saldos_prestamo donde los campos de union son pr_prestamos.(codigo_empresa,codigo_agencia,codigo_sub_aplicacion,numero_prestamo) en pr_saldos_prestamo.(codigo_empresa,codigo_agencia,codigo_sub_aplicacion,numero_prestamo)	MSALDOCUOTACAPITAL
	Si la operación es normal, cobro_judicial o inactiva (pr_prestamos.codigo_estado_cartera= A, B o C respectiva) el saldo de amortizacion será el valor contenido en pr_saldos_prestamo.valor cuyo pr_saldos_prestamos .codigo_tipo_saldo =20, si la operación esta en reserva(pr_prestamos.codigo_estado_cartera= D) entonces el saldo del amortización será el valor contenido en pr_saldos_prestamo.valor cuyo pr_saldos_prestamo.codigo_tipo_saldo = 36 Para obtener el saldo de amortización, se realizara un join entre pr_prestamos y pr_saldos_prestamo donde los campos de unión son pr_prestamos.(codigo_empresa,codigo_agencia,codigo_sub_aplicacion,numero_prestamo) en pr_saldos_prestamo.(codigo_empresa,codigo_agencia,codigo_sub_aplicacion,numero_prestamo)	MSALDOCUOTAAMORT

SISTEMA: CREDITO

	Si la operación es normal, cobro_judicial o inactiva (pr_prestamos.codigo_estado_cartera= A, B o C respectiva) el saldo de cuota interés será la suma de los valores contenidos en pr_saldos_prestamo.valor cuyo pr_saldos_prestamos .codigo_tipo_saldo =2 y 10, si la operacion esta en reserva(pr_prestamos.codigo_estado_cartera= D) entonces el saldo del cuota interés será la suma de los valores contenidos en pr_saldos_prestamo.valor cuyo pr_saldos_prestamo.codigo_tipo_saldo = 46 y 63 Para obtener el saldo de cuota interés, se realizara un join entre pr_prestamos y pr_saldos_prestamo donde los campos de unión son pr_prestamos.(codigo_empresa,codigo_agencia,codigo_sub_aplicacion,numero_prestamo) en pr_saldos_prestamo.(codigo_empresa,codigo_agencia,codigo_sub_aplicacion,numero_prestamo)	**MSALDOCUOTAINTERES**
	Si la operación es normal, cobro_judicial o inactiva (pr_prestamos.codigo_estado_cartera= A, B o C respectiva) el saldo de interés moratorio será el valor contenido en pr_saldos_prestamo.valor cuyo pr_saldos_prestamos .codigo_tipo_saldo =4, si la operación esta en reserva(pr_prestamos.codigo_estado_cartera= D) entonces el saldo del interés moratorio será el valor contenido en pr_saldos_prestamo.valor cuyo pr_saldos_prestamo.codigo_tipo_saldo = ? Para obtener el saldo de interés moratorio, se realizara un join entre pr_prestamos y pr_saldos_prestamo donde los campos de union son pr_prestamos.(codigo_empresa,codigo_agencia,codigo_sub_aplicacion,numero_prestamo) en pr_saldos_prestamo.(codigo_empresa,codigo_agencia,codigo_sub_aplicacion,numero_prestamo)	**MINTERESMORATORIO**
	Nota: Este proceso de extracción solo debe considerar operaciones activas pr_prestamos.estado = 1 no se debe considerar operaciones con pr_prestamos_cartera.codigo_estado_cartera = 99	

APÉNDICE D

MODELO ENTIDAD RELACIÓN FÍSICO


DWHP_TD_AGENCIA
DWHP_TD_EJECUTIVO
DWHP_TD_CLIENTE
DWHP_TD_SISTEMA
DWHP_TD_GRUPOINTERES
DMCR_TD_CALIFCLIENTE
DMCR_TD_TIPOCARTERA
DWHP_TD_MONEDA
DMCR_TD_FONDEO
DMCR_TD_ESTADOMORA
DMCR_TD_ACTIVIDAD
DMCR_TD_RUBRO
DMCR_TD_CLASE
DMCR_TD_SUBCLASE
DMCR_TH_CREDITO
DWHP_TD_AGENCIA
DMCR_TD_TIPOPRODUCTO
DMCR_TD_SUBPRODUCTO
DMCR_TD_PRODUCTO
DWHP_TD_SISTEMA
DMCR_TD_CALIFCREDITO
DWHP_TD_TIPOGARANTIA
DMCR_TD_NIVELAPROBACION
DMCR_TD_COMPCARTERA
DMCR_TD_SUBSECCION
DMCR_TD_ACTIVIDAD
DMCR_TD_SECCION

APÉNDICE E
PROGRAMAS DE EXTRACCIÓN Y CARGA DEL TANQUE

Como se explico en el capítulo 3, hemos utilizado programas de extracción para trasladar los datos, sin ninguna alteración, de los sistemas operacionales a una base de datos de trabajo, que llamaremos “tanque”. Este esquema nos permite transportar grandes cantidades de información sin sobrecargar, los sistemas fuentes con procesos adicionales. A la información contenida en esta base de datos se le aplican los procesos de transformación para el poblamiento del depósito de datos.

Para realizar esta labor se utilizó la herramienta gráfica “Data Transformation Services (DTS)” de Microsoft, como la herramienta que nos permitió la consolidación de datos de una gran variedad de fuentes y moverlos al depósito de datos, ya que es un utilitario para importar y exportar datos de múltiples fuentes heterogéneas. Esta fuera del alcance de este trabajo el brindar una explicación exhaustiva de cómo opera este producto, por lo remitimos al lector a la documentación técnica de SQL Server versión 7.0 o superior.

Por medio de la figura E.1 se visualiza este concepto, en donde gráficamente se construye un paquete de extracción y transformación de datos. Mediante la selección apropiada de los iconos de la barra de tareas y datos, el diseñador ubica y mueve los objetos al área de trabajo, estableciendo las tareas y la secuencia del flujo de trabajo.

A partir de una “conexión base”, se puede extraer y cargar información de fuentes tan variadas, como por ejemplo: un archivo texto, una hoja Excel hasta tablas de una base de datos, y aplicarle los algoritmos de transformación que permitan el poblamiento de bases de datos Oracle o SQL Server, todo esto agrupado bajo el concepto de paquete.

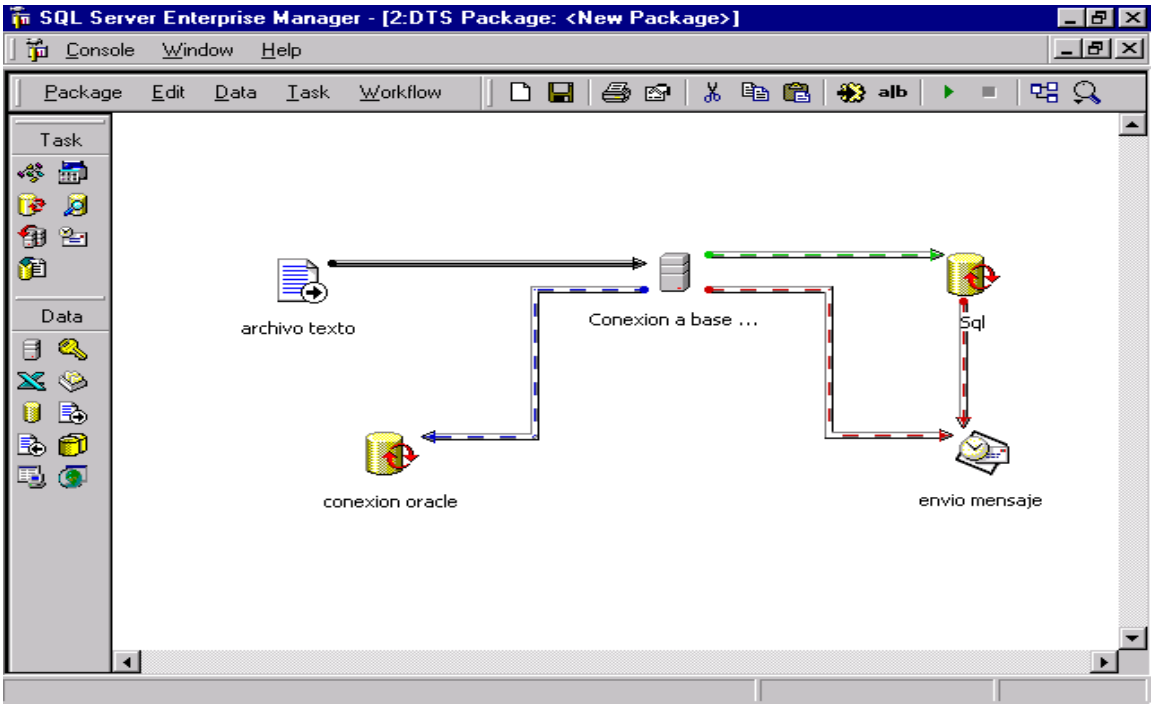


Figura E.1 Ejemplo de un diagrama de paquetes.

Se puede ver que la consola principal del diseñador de paquetes consiste de una barra de menú que contiene selecciones para las operaciones con los paquetes, tipos de fuentes de datos, tareas e ítem del flujo de trabajo. Cuenta también con un menú que contiene iconos para los diferentes tipos de fuentes de datos y tareas, además, de un espacio en el cual se ubican y conectan los objetos de manera gráfica, para establecer los diferentes flujos de trabajo.

APÉNDICE F

GUÍA PARA EL DESARROLLO DE UN DEPÓSITO DE DATOS

ARQUITECTURA DE LA PLANEACIÓN

El elemento más importante en la construcción de un depósito de datos es entender el tipo de decisiones que se requieren hacer y la información requerida para soportar estas decisiones, por lo tanto es necesario definir:

- Crear una visión.
- Definir alcances del proyecto.
- Definir metas y objetivos del proyecto.
- Estructurar el grupo de trabajo.
- Realizar el plan de trabajo.

ARQUITECTURA ACTUAL

Muchas metodologías proveen procesos que permiten definir, construir e integrar una arquitectura que soporte el esfuerzo del desarrollo de aplicaciones, incluyendo datos y tecnología. Es por ello se debe se debe tener un claro entendimiento de la arquitectura actual de la empresa:

- Arquitectura de aplicación: analizar los procesos de software que soportan los requerimientos funcionales del negocio.
- Arquitectura de datos: organizar las fuentes de información y el almacenamiento de los datos del negocio a lo largo de la empresa.
- Arquitectura de tecnología: conceptualizar la infraestructura tecnológica que permite que los datos y aplicaciones interactúen apropiadamente a lo largo de toda la empresa.

ARQUITECTURA DEL CICLO DE VIDA

Las principales fases en la construcción de un depósito de datos se centran en:

- Análisis:
  - Datos.
  - Aplicación.
  - Tecnología.

- Diseño:
  - Diagramas de paquetes de información.
  - Esquema estrella.
  - Modelo físico de la base de datos

Estas actividades conllevan a definir el dominio de la información, evaluar soluciones alternativas, esquematizar las especificaciones, mapear los requerimientos y especificaciones a la arquitectura.

<table>
<tr><th colspan="2">PLANTILLA PARA EL DESARROLLO DE UN DEPÓSITO DE DATOS</th></tr>
<tr><td>Nombre del Depósito de Datos</td><td></td></tr>
<tr><td>Departamento</td><td></td></tr>
<tr><td colspan="2">Visión</td></tr>
<tr><td>Metas</td><td>Objetivos</td></tr>
<tr><td></td><td></td></tr>
<tr><td rowspan="5">Principales entidades (medidas preliminares, dimensiones y categorías)</td><td></td></tr>
<tr><td></td></tr>
<tr><td></td></tr>
<tr><td></td></tr>
<tr><td></td></tr>
<tr><td>Revisado:

_____

Fecha:</td><td>Aprobado:

_____

Fecha:</td></tr>
</table>

PLANTILLA PARA LA DESCRIPCIÓN DE SISTEMAS		
Nombre del Sistema		
Departamento		
Función del negocio soportada		
Nombre del encargado del sistema		
Etapa del ciclo de vida:	( ) Planeado ( ) Desarrollo ( ) Producción	
Breve descripción del sistema		
Tecnología utilizada		
Hardware	**Software**	**Comunicaciones**
Entradas principales:		
Salidas principales:		
Modo de procesamiento: **( ) Lote ( ) En línea**	Horas pico de operación:	
Revisado: _____ Fecha:	Aprobado: _____ Fecha:	